PERICLES - Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics

[Digital Preservation]

# DELIVERABLE 4.4

# MODELLING CONTEXTUALISED SEMANTICS


***GRANT AGREEMENT: 601138***

***SCHEME FP7 ICT 2011.4.3***

***Start date of project: 1 February 2013***

***Duration: 48 months***

| Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013) | | |
|---|---|---|
| **Dissemination level** | | |
| **PU** | PUBLIC | X |
| **PP** | Restricted to other PROGRAMME PARTICIPANTS (including the Commission Services) | |
| **RE** | RESTRICTED to a group specified by the consortium (including the Commission Services) | |
| **CO** | CONFIDENTIAL only for members of the consortium (including the Commission Services) | |

# Revision History

| V # | Date | Description / Reason of change | Author |
|---|---|---|---|
| V0.1 | 22.02.16 | Outline and first draft. | CERTH |
| V0.2 | 25.03.16 | Additions to Chapters 3, 4, 5, 6. | CERTH, HB |
| V0.3 | 15.04.16 | Additions to Chapter 6, refined Chapters 4 and 6. | KCL, CERTH |
| V0.4 | 29.04.16 | Preparing for 1st draft. | CERTH |
| V0.5 | 10.05.16 | 1st draft submitted to internal reviewers. | CERTH, HB |
| V0.6 | 23.05.16 | Integrated suggestions by internal reviewers, working on Chapters 1, 2, 6. | CERTH, HB |
| V1.0 | 25.05.16 | Final version to be submitted to reviewers. | CERTH, HB |

# Authors and Contributors

## *Authors*

| Partner | Name |
|---|---|
| CERTH | E. Kontopoulos, M. Riga, P. Mitzias, S. Andreadis, T. Stavropoulos, K. Konstantinidis, A. Maronidis, A. Karakostas, S. Tachos, V. Kaltsa, M. Tsagiopoulou |
| HB | S. Darányi, P. Wittek |
| KCL | A. Gill, E. Tonkin |

## *Contributors*

| Partner | Name |
|---|---|
| KCL | S. Waddington, C. Sauter |
| ULIV | F. Corubolo |

## *Reviewers*

| Partner | Name |
|---|---|
| XRCE | N. Lagos |
| UGOE | J. Biermann |

# Table of Contents

# Glossary

| Abbreviation / Acronym | Meaning |
|---|---|
| **A&M** | Art & Media |
| **ANN** | Artificial Neural Network |
| **API** | Application Program Interface |
| **BDA** | Born-Digital Archives |
| **BMU** | Best Matching Units |
| **CB** | Computer-based |
| **CH** | Cultural Heritage |
| **CM** | Classical Mechanisms |
| **DEM** | Digital Ecosystem Model |
| **DL** | Description Logics |
| **DO** | Digital Object |
| **DoW** | Description of Work |
| **DP** | Digital Preservation |
| **DVA** | Digital-Video Artworks |
| **ESOM** | Evolving Self-Organizing Maps |
| **HCA** | Hierarchical Cluster Analysis |
| **IR** | Information Retrieval |
| **ISAD(G)** | General International Standard Archival Description |
| **JSON** | JavaScript Object Notation |
| **LRM** | Linked Resource Model |
| **LSA** | Latent Semantic Analysis |
| **LTDP** | Long-term Digital Preservation |
| **ML** | Machine Learning |
| **MM** | Mixed-media |
| **OAIS** | Open Archival Information System |
| **ODP** | Ontology Design Pattern |
| **ORSD** | Ontology Requirements Specification Document |

| | |
|---|---|
| **OS** | Operating System |
| **OWL** | Web Ontology Language |
| **PCA** | Principal Component Analysis |
| **QA** | Quality Assurance |
| **RIX** | Random IndeXing |
| **RDF** | Resource Description Framework |
| **SB** | Software-based |
| **SBA** | Software-Based Artworks |
| **SOM** | Self-Organizing Maps |
| **Somoclu** | Self-Organizing Maps Over a CLUster |
| **SPIN** | SPARQL Inferencing Notation |
| **VSM** | Vector Space Model |

# 1. Executive Summary

The current deliverable summarises the work conducted within task T4.4 of WP4, presenting our proposed models for semantically representing digital content and its respective context – the latter refers to any information coming from the environment of the digital object (DO) that offers a better insight into the object's status, its interrelationships with other content items and information about the object's context of use. Within PERICLES, we refer to the content semantics enriched with the contextual perspective as "contextualised semantics". The deliverable presents two complementary modelling approaches, based respectively on (a) ontologies and logics, and, (b) multivariate statistics.

Additionally, D4.4 also studies semantic change and discusses our proposed methodologies for its detection, measurement and interpretation, presenting a set of relevant experiments with different aspects of partner data aiming at visualising and finding solutions to semantic drifts.

More specifically, the main contributions of the deliverable are the following:

- Extensive and up-to-date state-of-the-art surveys on semantically representing content and context via ontologies.
- Novel, highly modular and extensible ontology-based models for semantically representing digital content, context and use-context (i.e. context of use), based on the Linked Resource Model (LRM), a core PERICLES output.
- An inference layer on-top of the developed models for taking advantage of context representation and contextualised content semantics, facilitating automated reasoning and handling of various inconsistencies.
- A novel method combining semantic fields from linguistics, multivariate statistics, and the concept of fields in classical mechanics to study context-dependent evolving semantics as a vector field.
- A technology to detect, measure and interpret semantic drifts in scalable and dynamically evolving text and image metadata collections.
- A thorough background survey of all notions relevant to semantic change and the overall phenomenon of evolving semantics, along with an attempt to disambiguate the various respective terms found in this rapidly growing area of research.
- Three directions of novel research in the area of semantic change for theory verification by experiments: (a) a field approach to evolving semantics, dealing with textual content and indexing terminology change, (b) a study of semantic change under an ontology evolution perspective, investigating changes occurring in ontology models, and, (c) a study of community change in social media.
- A fourth very interesting "guest" line of analytical work on topic shifts which is not included in the methodological spectrum of PERICLES by affiliated partners outside the project's consortium.
- Respective open-source implementations and experimental results that validate all the above.

All these contributions are tightly interlinked with the other PERICLES work packages: WP2 supplies the use cases and sample datasets for validating our proposed approaches, WP3 provides the models (LRM and Digital Ecosystem models) that form the basis for our semantic representations of content and context, WP5 provides the practical application of the technologies developed to preservation processes, while the tools and algorithms presented in this deliverable can be deployed in combination with test scenarios, which will be part of the WP6 test beds.

# 2. Introduction & Rationale

PERICLES relies heavily on semantically representing digital content (i.e. resources) along with its environment (i.e. the digital ecosystem the resources reside in) and context of use. Within PERICLES, the content semantics enriched with the contextual perspective is referred to as "**contextualised semantics**", although in literature this term typically refers to linguistics and the phenomenon of words changing their meaning depending on the coexistence of other words in their immediate environment in a document.

Thus, summarising the work of task T4.4, this deliverable **presents our proposed models for semantically representing concepts and their context**. The semantic representation of content refers to the notions, meanings, topics and themes encompassed by the object typically represented via structures like e.g. ontologies and semantic networks, in order to capture the conceptual representations of terms. Semantically representing context, on the other hand, refers to modelling the relationships of content items to one another and to pertinent aspects in their environment, along with information relevant to the context of use of the items.

Additionally, D4.4 also studies **semantic change**, which is a growing area of research that observes and measures the phenomenon of modifications in the meaning of concepts within knowledge representation models. Since semantic change can have drastic consequences on accessing digital content, this deliverable discusses our **proposed methodologies for detecting, measuring and interpreting semantic change** (by means of conceptual and contextual semantics), along with relevant experiments with different aspects of partner data and beyond, aiming at visualising and finding solutions to semantic drifts.

Finally, this document **explores how all of these investigations relate to the other relevant research activities** within PERICLES.

## 2.1. What to expect from this Document

This deliverable constitutes the output of the activities taking place within T4.4 and discusses the following topics:

- **Semantic representation of content**: The document goes through our adopted approaches for representing semantic content, providing a detailed account of the specification, formalisation and implementation of the proposed models for the two use case domains within PERICLES.
- **Context modelling and contextualised content semantics**: The deliverable presents our proposed approaches for semantically representing context and use-context (i.e. context of use) of resources residing in digital ecosystems, along with a proposed approach for taking advantage of contextualised content semantics, in order to infer variations in meaning and interpretation according to the context in which content is viewed.
- **Statistical context modelling and evolving semantics in a vector field**: In this context, the document presents a novel methodology to treat context-dependent, evolving semantic content as a vector field by a combination of ideas from linguistics, statistics and classical mechanics.
- **Semantic change and evolving semantics**: Finally, based to a great extent on the proposed representations for content and context, the deliverable studies semantic change along with the overall phenomenon of evolving semantics and presents our lines of investigation in this area.
- Respective **open-source implementations and experimental results** that validate all the above.

## 2.2. Relation to other Work Packages

The work presented in this deliverable is strongly linked to other PERICLES WPs as follows:

- The models proposed here for semantically representing content and context in the two domains (Space Science and Art & Media) are heavily based on the domain ontologies developed within **WP2**. Nevertheless, this relationship with the domain ontologies is bidirectional, in the sense that work conducted within WP4 also feeds into the domain ontologies, revealing additional constructs and representations to be adopted by the latter.
- Similarly, a significant part of our investigations on semantic change are based on the models developed within **WP2**. Also, a portion of the relevant methodologies have been deployed on Tate's online artwork repositories.
- Furthermore, our semantic change investigations also feed into **WP3** (Digital Ecosystem Model) and **WP5** (QA and appraisal tools). The detection and measurement of semantic drift as one type of change affecting Digital Preservation (DP) systems connects on the one hand to **WP3**, creating a passage between ontology updates and statistical updates as two approaches to the same entity in study. On the other hand, semantic drift quantification paves the way for an at-risk terminology alert service in **WP5** and its pilot implementation in WP6.
- Finally, the developed models that are presented in this deliverable serve as the underlying knowledge bases of the **WP6** test beds running within the integration framework and, thus, constitute an integral part.

## 2.3. Relation to other WP4 Tasks

Besides the interconnections with other WPs, the work presented here is tightly linked to the other WP4 tasks as well. Thus, parts of the underlying modelling outputs from **T4.4** are exploited by the extraction and encapsulation activities from **T4.1** and **T4.2** and by the analysis activities in **T4.3**. Similarly, the upcoming D4.5 describing the interpretation and reasoning activities in **T4.5** will also be based on the models described in this deliverable. More specifically:

- The work conducted within **T4.4** is the entry point to **T4.3** "*Semantic content and use-context analysis*", looking at text and image content analytic methods from a context-dependent perspective. Here, the theoretical underpinnings are clarified, to be taken over by **T4.3** where experimental evidence proves their scalable usability.
- **T4.1** (PET) and **T4.2** (PET2LRM) feed into our models for semantically representing context and use-context, as described in D4.3 [PERICLES D4.3, 2016].
- **T4.5** "*Contextualised content interpretation*" is directly connected with the field approach to evolving semantics in D4.4, because the analytic effort to identify quantum-likeness in LTDP-relevant system behaviour departs from the same physical metaphor. Comparing similarity between DOs and their features to a force, "conceptual mass" and "conceptual energy" as generative components of this statistical force are considered, crossing the no man's land between classical and quantum mechanics. On the other hand, as evolving vector spaces correspond to evolving graphs, there is a direct link to ontology evolution and dynamic reasoning based on the LRM and its domain-specific spin offs.

## 2.4. Document Structure

The structure of the rest of this document is as follows:

- **Chapter 3 - Ontology-based Representation of Content and Context**: This chapter presents our proposed models for semantically representing content and context in the two use case domains of the project. The representations are based on the domain ontologies developed

within the project in OWL (Web Ontology Language) and Topic Maps that are built on top of the Linked Resource Model (LRM), a core PERICLES outcome. The semantic representation of use-context is based on the LRM Dependency construct, taking advantage of its encompassed specification and intention. Finally, the chapter also discusses our adopted contextualised semantics approach, according to which, we are proposing an additional inference layer on-top of the developed models for handling context-related inconsistencies. This validation layer uses SPIN, the SPARQL Inferencing Notation, a well-known notation for representing SPARQL rules and constraints on models, and for performing queries on RDF graphs.

- **Chapter 4 - Statistical Context Modelling and Contextualised Content Semantics**: Based on multivariate statistics for scalability, tool and methodology testing, this chapter presents an approach different from the one in Chapter 3, working with ontology-based representations. At the same time it paves the way for theory development key in chapter 5 and related research in 2016.
- **Chapter 5 - Semantic Change and Evolving Semantics**: This chapter discusses semantic change and the overall phenomenon of evolving semantics and presents a thorough background survey of all relevant notions, in an attempt to disambiguate the various respective terms found in this rapidly growing area of research. Three directions of novel semantic change research are presented in chapter 5: (a) a field approach to evolving semantics, dealing with textual content and indexing terminology change, (b) a study of semantic change under an ontology evolution perspective, investigating changes occurring in ontology models, and, (c) a study of community change in social media. A fourth “guest” line of analytical work on topic shifts by affiliated partners outside the project’s consortium is also presented in this chapter, which is not included in the methodological spectrum of PERICLES. The developed software tools for these investigations are publicly available along with the respective results and datasets.
- **Chapter 6 - Conclusions & Future Work**: Finally, the deliverable concludes with some final remarks and an account of potentially interesting directions for future work, with regards to each of the key topics discussed in the previous chapters.

# 3. Ontology-based Representation of Content and Context

This chapter describes our ontology-based approaches for semantically representing content and context. After the scope of the proposed semantic models is briefly given, the chapter provides a thorough account of the state-of-the-art in ontology-based representation of content semantics, followed by a description of the adopted semantic models, focusing on their respective specification, formalisation and implementation. Then, the chapter proceeds with discussing our proposed approaches: (a) for semantically representing contextual information aspects, and (b) for taking advantage of this information in order to infer different variations in meaning and interpretation. Specifically for point (a), the chapter features a state-of-the-art survey on ontology-based schemes for modelling contextual information focusing on DP-related approaches, followed by our suggested modelling methodologies for representing context and use-context. Details on the respective implementations in OWL are also given. For point (b), the chapter revolves around "contextualised content semantics" (i.e. a term used for referring to variations in the meaning and interpretation of units of content that arise according to the context in which that content is viewed), and introduces our proposed methodology for using context representation towards creating an additional inference layer on-top of the developed models.

## 3.1. Content Modelling

### *3.1.1. Scope of the PERICLES Semantic Content Models*

The project's domain ontologies are the primary means for **semantically representing content** in PERICLES and are based on the LRM [PERICLES D3.2, 2014; PERICLES D3.3, 2015] and the Digital Ecosystem Model (DEM) [PERICLES D5.2, 2015]. The aim of the domain ontologies is not to exhaustively model the respective case study domains, but to **serve as the foundations for deploying the novel DP methodologies** proposed within PERICLES. Thus, the Art & Media (A&M) domain ontologies are primarily aimed at modelling DP-related risks in the three subdomains (digital video art, software-based art and born-digital archives) and are expected to facilitate curators in modelling, projecting and tackling risks throughout several phases of the whole DP process. The Space Science domain ontology, on the other hand, encompasses concepts and details that are deemed relevant for capturing the processes and data flows for the SOLAR experiment inputs and outputs and establishes links to the scientific data and its provenance. All in all, the domain ontologies can also be seen as a means to increase the understanding of the case studies, as they make links and dependencies between concepts explicit. More details about the scope of each domain ontology are given in D2.3.2 [PERICLES D2.3.2, 2015].

Currently, the semantic domain models are considered sufficient in breadth and depth for their intended purposes, but slight refinements and fine-tuning are expected to take place on a by-need basis during the final months of the project. All in all, the proposed models are flexible enough, so that if future adopters wish to add more fine-grained modelling elements, depending on certain application requirements, this can be easily achieved via extending the models appropriately.

### *3.1.2. State-of-the-Art in Ontology-based Content Modelling*

This section presents the state-of-the-art in ontology-based representation of content, pertinent to the PERICLES domains. Thus, a brief account of most established multimedia and cultural heritage ontologies is given, followed by a discussion of other relevant vocabularies.

### *MULTIMEDIA ONTOLOGIES*

As stated above, the purpose of the PERICLES ontologies is not to exhaustively model the respective case study domains. Therefore, although specified in the Description of Work (DoW), we didn't consider using and/or extending existing multimedia ontologies, since the latter are focused on a detailed description of multimedia and audiovisual content. Nevertheless, a state-of-the-art survey was carried out, in order to consider notions and concepts we could potentially include in our developed models.

All in all, multimedia ontologies are distinguished in MPEG-7-compliant and noncompliant ones. **MPEG-7** has been an established multimedia content description standard since 2001, when it achieved an ISO/IEC 15398 status. Unlike other standards, like e.g. MPEG-1, MPEG-2 and MPEG-4, that deal with the actual encoding of moving pictures and audio, MPEG-7 provides complementary functionality and represents information about the content, assisting in the fast and efficient search for digital multimedia resources. MPEG-7 is formally called "**Multimedia Content Description Interface**" and uses XML for representing metadata, while applications that benefit from the MPEG-7 standard range across content management, organization, navigation and automated processing.

The following is an overview of multimedia ontologies that conform to the MPEG-7 standard. More information is available in related survey papers, like e.g. [Dasiopoulou et al., 2010].

**Harmony (2001)**

This is the first attempt to formally represent an MPEG-7-based ontology [Hunter, 2001]. The ontological representation of the model is an accurate translation of the initial MPEG-7 definitions. Consequently, the ambiguities present in the original MPEG-7 specification are propagated in the ontology model as well, resulting in implications on the conceptual clarity and subsequent management of the represented descriptions.

**AceMedia (2004)**

Within the AceMedia project[1] two ontologies were developed: the Multimedia Structure Ontology (MSO) and the Visual Descriptor Ontology (VDO) [Bloehdorn et al., 2004; Simou et al., 2005]. Specifically, MSO represents the structural description tools from the Multimedia Description Schemes (MDS) of the MPEG-7 specification, while VDO handles the representation of the visual part. The design of the two ontologies follows the principles of the Harmony ontology described previously, resulting in similar semantic ambiguity issues, which, however, are partially alleviated by a better level of granularity and the presence of two diverse ontologies instead of a single ontology that allows for a more modular engineering approach.

**Rhizomik (2005)**

The Rhizomik ontology [Garcia & Celma, 2005] represents an attempt to perform a fully automatic translation of the complete MPEG-7 Schema to OWL; the result is an OWL DL ontology, covering all elements of the entire MPEG-7 specification. Similarly to the Harmony and AceMedia ontologies, Rhizomik preserves the flexibility of the MPEG-7 specifications. As a result, all ambiguities present in MPEG-7 are retained. However, the model is not detailed enough and defines a rather coarse conceptualisation when linked to existing domain ontologies.

**SmartWeb (2006)**

The SmartWeb ontology was proposed for supporting the annotation of multimedia content [Oberle et al., 2007; Vembue et al., 2006]. The specific approach handles the representation of the structural,

[1] AceMedia project: http://www.acemedia.org

localisation, media and low-level MPEG-7 descriptions and realises a meta-modelling ontological framework that allows to formally model the MPEG-7 descriptions and export them into OWL and RDFS. Additionally, linking with domain-specific ontologies is achieved via a specific infrastructure that aligns the developed set of ontologies. On the downside, although the modelling perspectives followed by the SmartWeb ontology are closer to the original MPEG-7 Schemas, additional semantic ambiguities are introduced, especially in the case of recursive content decomposition.

**Boemie (2007)**

In an attempt to capture the semantics of the MPEG-7 structural descriptions, two OWL-DL ontologies have been proposed in the context of the BOEMIE project[2] [Dasiopoulou et al., 2007]: Multimedia Content Ontology (MCO) and Multimedia Descriptors Ontology (MDO). Instead of following a strict translation from the MPEG-7 specifications, MCO re-engineers the MPEG-7 structural and localisation descriptions in order to axiomatise the intended meaning and introduces an array of new features. Linking with domain specific ontologies is achieved through a pair of generic properties that capture the relation between a content/segment instance and the depicted semantics, and the relation between a content/segment instance and its extracted low-level features.

**M-OWL (2006)**

M-OWL [Harit et al., 2006] is an extension to the Web Ontology Language (OWL) that makes use of the standard descriptors provided by MPEG-7, but also defines additional descriptors using the MPEG-7 Description Definition Language (DDL). M-OWL definitions for media related observations incorporate the flexibility of making use of MPEG-7 descriptions, as well as other markups like RuleML and MathML. In M-OWL, it is possible to associate different types of media features in different media formats and at different levels of abstraction with the concepts in a closed domain. Finally, M-OWL supports an abductive reasoning framework using Bayesian networks that is robust against imperfect observations of media data.

**DS-MIRF (2007)**

A further approach is represented by the DS-MIRF framework [Tsinaraki et al., 2007; Tsinaraki & Christodoulakis, 2007], which constitutes a manual translation of the complete MPEG-7 MDS into an OWL-DL ontology. As in the case of Rhizomik, a one-to-one translation has been followed taking into account all elements appearing in the respective MPEG-7 description tools. However, the DS-MIRF approach attempts to make explicit the implicit notions of the initial schemas, resulting in improved clarity of the translation semantics. This is an advantage that Rhizomik's automated transformation cannot offer.

**COMM (2007)**

COMM (Core Ontology for MultiMedia) is a well-founded multimedia ontology framework, providing a comprehensive capability to annotate non-textual media [Arndt et al., 2007]. The ontology was built by re-engineering and formalizing MPEG-7. In order to support conceptual clarity and soundness as well as extensibility towards new annotation requirements, COMM builds on a popular foundational ontology, Descriptive Ontology for Linguistic and Cognitive Engineering - DOLCE [Gangemi et al., 2002], and is divided up into modules. All in all, COMM is based on a careful analysis of the requirements underlying the semantic representation of media objects and goes beyond the capabilities of most semantic multimedia ontologies.

[2] BOEMIE project: http://www.boemie.org/

**OMR (2012)**

The Ontology for Media Resources (OMR)[3] is a W3C Recommendation and constitutes a core vocabulary for cross-community data integration of information related to multimedia resources available on the Web. In essence, OMR is a lightweight ontology that attempts to bridge the various available descriptions of media resources, providing a core set of descriptive properties. However, besides merely providing a core vocabulary, OMR also serves as a mapping schema to a set of popular metadata formats describing media resources published on the Web. In reality, however, this is not easily achieved, since each format possibly covers different extensions of the same term and mapping back and forth between properties from different schemas via OMR may lead to loss in semantics.

**Custom Multimedia Ontologies**

Besides MPEG-7-compliant multimedia ontologies, there also exist approaches that adopt customized ad hoc modelling choices that are proposed within specific applications. Below are some of the most representative paradigms.

- In [Hudelot & Thonnat, 2003; Maillot & Thonnat, 2005] a visual ontology is proposed that provides qualitative descriptions with respect to colour, texture and spatial aspects of the characterised content.
- Similar qualitative visual descriptors have been deployed in the Breast Cancer Imaging Ontology - BCIO [Hu et al., 2003].
- In [Goodall et al., 2003] an ontology for representing museum multimedia resources is proposed, accompanied by a graphical concept browser interface for navigating through the ontology and displaying the various content types to the appropriate viewers.
- In [Hollink & Worring, 2005] a "visual ontology" is proposed that combines WordNet and MPEG-7 descriptions for representing the various visual attributes of objects, such as shape, colour, visibility, etc.

As it is easily understood, the above approaches don't invest much in interoperability, contrary to MPEG-7 based multimedia ontologies. However, although they all share a common vision, they also introduce several conceptual differences regarding the modelling of content semantics as well as the linking with domain ontologies.

### *CULTURAL HERITAGE ONTOLOGIES*

The **CIDOC Conceptual Reference Model** (**CIDOC CRM**), acknowledged as an ISO Standard (21127:2006) is the most dominant ontology for knowledge representation in cultural heritage (CH) and museum documentation [Doerr, M., 2005]. Its basic aim is to facilitate the integration and interchange of heterogeneous CH information among distributed digital sources. The semantic definitions and the formal structure provided by the ontology establish an interoperable global resource and promote a shared understanding in the domain. The key concepts covered by CIDOC CRM are: (a) *persistent items* that represent items with a persistent identity; they can be either physical (people, things) or conceptual (ideas, concepts, products) entities, and, (b) *temporal entities* representing all phenomena that occur over a limited or extended time. Activities are essential temporal entities, therefore CIDOC incorporates some activity types that may be related to works of art or other cultural creations, like e.g. the *creation*, the *acquisition* and the *modification* of a work.

The **FRBRoo**[4] (**Functional Requirements for Bibliographic Records-Object Oriented**) is a formal ontology intended to capture and represent the underlying semantics of bibliographic information

[3] W3C Ontology for Media Resources 1.0: http://www.w3.org/TR/2012/REC-mediaont-10-20120209/
[4] http://www.cidoc-crm.org/frbr_inro.html

and to facilitate the integration, mediation and interchange of bibliographic and museum information. The underlying FRBR model was originally designed as an entity-relationship model by the International Federation of Library Associations and Institutions (IFLA) during the period 1991-1997, and was published in 1998. Quite independently, the CIDOC CRM model was being developed from 1996 under the auspices of the ICOM-CIDOC (International Council for Museums – International Committee on Documentation) Documentation Standards Working Group. The idea that both the library and museum communities might benefit from harmonising the two models was first expressed in 2000 and grew up in the following years. Eventually, it led to the formation, in 2003, of the International Working Group on FRBR/CIDOC CRM Harmonisation, that brings together representatives from both communities with the common goals of: a) expressing the IFLA FRBR model with the concepts, tools, mechanisms, and notation conventions provided by the CIDOC CRM, and: b) aligning (possibly even merging) the two object-oriented models with the aim to contribute to the solution of the problem of semantic interoperability between the documentation structures used for library and museum information.

**Europeana**[5] is a multilingual digital library that was first introduced in 2008 and is aimed at promoting the collaboration between museums, libraries and collections and facilitating user access to an integrated content for European cultural and scientific heritage. The library currently encompasses descriptions for more than 20 million digital objects. Europeana collects contextual information and metadata about the items provided by individual CH institutions. Its main novelty is that, rather than attempting to produce one unified ontology, Europeana tries to establish alignments between local vocabularies (used to annotate the original data) and more general pivot vocabularies [Charles & Isaac, 2015]. With semantically enriched metadata, Europeana establishes associations in a common context and automatically suggests materials to the user that are related to a particular retrieved object.

The **Getty Vocabularies**[6], developed by the Getty Research Institute (GRI), provide structured terminology for works of art, architecture, material culture, as well as artists, architects and geographic locations. The vocabularies, whose development started in the late 70s, are currently four: (a) the *Art & Architecture Thesaurus* (*AAT*) that describes concepts related to artworks, architecture and archeology, (b) the *Getty Thesaurus of Geographic Names* (*TGN*), including names and descriptions of modern or historical places of CH importance, (c) the *Cultural Objects Name Authority* (*CONA*), a structured vocabulary that contains names and records of cultural works, such as built (architecture) and movable (sculpture, paintings, photographs, etc.) work, and, (d) the *Union List of Artist Names* (*ULAN*), which is a thesaurus of artist names and information. The Getty vocabularies are freely available for online use and research.

**CRM Digital** (**CRMdig**) is an extension of the CIDOC-CRM ontology that supports "provenance" metadata, meaning information about the steps, the methods and the relations of digitization products [Doerr & Theodoridou, 2011]. The main objective of the ontology is to describe the origins and derivations of digital items and to define all (historical) interrelationships among different versions of a digital item. It also describes the devices that participate in the measurement or digitization and makes it possible to follow the history of individual devices, track factors of possible distortion of results and answer complex queries regarding their status. The main classes of CRM*dig* comprise the hierarchy below:

- *Digital Object*, which describes any immaterial item that can be represented as a bit sequence, like audio and video items, images, software and e-texts.

[5] http://www.europeana.eu/portal
[6] http://www.getty.edu/research/tools/vocabularies

- *Digital Machine Event*, which is a generic notion that describes events performed on physical digital devices throughout a human activity and result in the creation of a new instance of a Digital Object.

### OTHER RELEVANT VOCABULARIES

The **Dublin Core** (**DC**) **Metadata Element Set**[7] is a set of vocabulary terms that can be used to describe resources, i.e. from web resources (video, web pages, etc.) to physical resources (books, artworks, etc.). The DC Metadata Initiative began in 1995 and, up to now, two forms of DC vocabularies exist. These are:

- the *Simple Dublin Core*, which is the original set that consists of 15 metadata elements (*title*, *creator*, *subject*, *description*, *publisher*, *contributor*, *date*, *type*, *format*, *identifier*, *source*, *language*, *relations*, *coverage* and *rights*)[8], and
- the *Qualified Dublin Core*, which enriches the specificity of metadata with three additional elements (*audience*, *provenance* and *rights holder*) as well as a number of so called qualifiers that refine the semantics of elements in a narrower or more specific meaning.

Concepts and semantics are designed to be independent and equally applicable in a variety of domains and contexts, as well as in our domain of interest.

### ONTOLOGY DESIGN PATTERNS AND MAPPING TO LRM CONCEPTS

**Ontology Design Patterns** (**ODP**s) are used in many fields as modelling "templates" or abstract descriptions encoding best practices of some field[9]. Several ODPs could be incorporated and reused while creating the Art & Media (A&M) domain ontologies. However, the LRM already includes corresponding constructs and covers all desired modeling areas. A list of existing ODPs and their field of interest are the following (see also summary information in Table 3-1):

- **Representing mereology (part-of relationships)** - *Componency*[10] and *PartOf*[11] are patterns that represent entities and their parts using properties such as `hasComponent` and `hasPart`. The LRM counterparts are the object properties `lrm:hasPart` and `lrm:partOf`.
- **Representing realisation of information objects (i.e. DOs)** - Pattern *Information Realization*[12] allows distinguishing information objects from their concrete realisations with the use of property `isRealizedBy`. Respectively, the LRM has properties `lrm:realizes` and `lrm:realizedAs`.
- **Representing aggregated objects** - For pattern *SimpleOrAggregated*[13], several objects gathered in another object acting as a whole are represented with classes `SimpleObject` and `AggregatedObject` and property `hasAggregatedMember`. Furthermore, patterns named *Collection*[14] and *Collection Entity*[15] aim at representing collections of objects/resources via the class `Collection` and the properties `isMemberOf/hasMember`. For the aforementioned concepts, the LRM incorporates classes `lrm:AggregatedResource` and properties `lrm:hasPart`, `lrm:partOf`.

[7] http://dublincore.org/
[8] A full dataset can be found at http://dublincore.org/documents/dcmi-terms/
[9] http://ontologydesignpatterns.org/wiki/Odp:WhatIsAPattern
[10] http://ontologydesignpatterns.org/wiki/Submissions:Componency
[11] http://ontologydesignpatterns.org/wiki/Submissions:PartOf
[12] http://ontologydesignpatterns.org/wiki/Submissions:Information_realization
[13] http://ontologydesignpatterns.org/wiki/Submissions:SimpleOrAggregated
[14] http://ontologydesignpatterns.org/wiki/Submissions:Collection
[15] http://ontologydesignpatterns.org/wiki/Submissions:CollectionEntity

- **Representing descriptions** - The LRM counterpart of ODP *Description*'s[16] properties `defines` and `isDefinedBy` are respectively properties `lrm:describes` and `lrm:describedBy`.
- **Representing activities** - ODP *Activity Reasoning*[17] provides a generic pattern for modelling the common core of activities in different domains. Similarly, `lrm:Activity` with extra classification and corresponding properties covers this modelling area.
- **Representing time intervals** - *Time Interval*[18] is a design pattern to represent time intervals, using properties `hasIntervalDate`, `hasIntervalStartDate` and `hasIntervalEndDate`. Properties `lrm:starting` and `lrm:ending` may also serve the same cause.
- **Representing general types of entities** - *Types of entities*[19] is a pattern that tries to categorize the most general types of things in the domain with classes `Abstract`, `Object`, `Event` and `Quality`. On the contrary, the LRM suggests a slightly different classification, with top level entities `lrm:AbstractResource`, `lrm:ConcreteResource` and `lrm:AggregatedResource`.
- **Representing sequence** - The *Sequence*[20] ODP suggests a way to represent sequence schemas with properties `precedes` and `follows`, that is useful for time lines, event sequences, versions. Related properties exist in LRM, which are `lrm:preceding` and `lrm:following`.
- **Representing location** - The *Place*[21] pattern is a simple structure for defining the location of a certain thing. The LRM incorporates the class `lrm:Location` and its connected properties, to define information about the location of a concrete resource.

**Table 3-1**. Analogy between existing ODPs and LRM constructs.

| **Field of representation** | **ODP constructs** | **LRM constructs** |
|---|---|---|
| Mereology (part-of relationships) | Componency: `hasComponent`, `isComponentOf`<br>PartOf: `hasPart`, `isPartOf` | `lrm:hasPart`, `lrm:partOf` |
| Realization of information objects | Information Realization: `isRealizedBy` | `lrm:realizes`, `lrm:realizedAs` |
| Aggregated objects | Simple Or Aggregated: `SimpleObject`, `AggregatedObject`, `hasAggregatedMember`<br>Collection, Collection Entity: `Collection`, `isMemberOf`, `hasMember` | `lrm:AggregatedResource`, `lrm:hasPart`, `lrm:partOf` |
| Description of objects | Description: `defines`, `isDefinedBy` | `lrm:describes`, `lrm:describedBy` |

[16] http://ontologydesignpatterns.org/wiki/Submissions:Description
[17] http://ontologydesignpatterns.org/wiki/Submissions:An_Ontology_Design_Pattern_for_Activity_Reasoning
[18] http://ontologydesignpatterns.org/wiki/Submissions:TimeInterval
[19] http://ontologydesignpatterns.org/wiki/Submissions:Types_of_entities
[20] http://ontologydesignpatterns.org/wiki/Submissions:Sequence
[21] http://ontologydesignpatterns.org/wiki/Submissions:Place

| Field of representation | ODP constructs | LRM constructs |
|---|---|---|
| Activities | Activity Reasoning: `Activity`, `Requirement`, `foaf:Agent`, `xsd:duration`, etc. | `lrm:Activity` with extra classification and corresponding properties |
| Time intervals | Time Interval: `hasIntervalDate`, `hasIntervalStartDate`, `hasIntervalEndDate` | `lrm:starting`, `lrm:ending` |
| General types of entities | Types of entities: `Abstract`, `Object`, `Event`, `Quality` | `lrm:AbstractResource`, `lrm:ConcreteResource`, `lrm:AggregatedResource` |
| Sequence | Sequence: `precedes`, `follows` | `lrm:preceding`, `lrm:following` |
| Place | Place: `Place`, `isLocationOf` | `lrm:Location`, `lrm:location` |

## *3.1.3. Semantic Representation of Content*

This subsection presents our proposed scheme for semantically representing content, focusing on the design decisions taken during its specification and formalisation, followed by details of its implementation.

### *SPECIFICATION*

Contrary to the Space Science domain, where the semantic representation is undertaken by a single ontology, for the A&M domain we have developed three specific domain-related ontologies [PERICLES D2.3.2, 2015], which are: (a) the Digital-Video Artwork (DVA), (b) the Software-Based Artwork (SBA), and (c) the Born-Digital Archives (BDA). Several key challenges have been defined within each of these subdomains and corresponding ontologies have been developed, which do not attempt to exhaustively model the respective subdomains, but are primarily aimed at modelling specific DP-related risks that demonstrate an interesting range of DP challenges in the domain of interest. Specifically, regarding DVA, the focus is on the **consistent playback of digital video files**, with respect to the technical or conceptual characteristics of the corresponding digital components. Concerning SBA, the focus is on the **assessment of risks for newly acquired artworks**, regarding their technical dependencies and the ability to be displayed properly, consistently and accurately. Finally, within the BDA context, the focus is on the need of being able to **access and maintain digital documents as they were initially meant to be**, with all technical, aesthetical, permission characteristics that they are attached to.

Despite the fact that the three A&M subdomains are quite distinct from each other, the following common notions were adopted during the design of all three ontologies (DVA, SBA and BDA):

**Abstract** (`lrm:AbstractResource`), **Concrete** (`lrm:ConcreteResource`) and **Aggregated Resource** (`lrm:AggregatedResource`) - represent the most high-level distinction between resources existing in the domain of interest. An abstract resource is a concept of an entity that may be implemented (`lrm:realizedAs`) in one or more concrete resources. If the realisation of an

entity contains more than one resources, then this is represented via an aggregated resource and the different parts are connected with the aggregated instantiation via the property `lrm:hasPart`.

**Activity** (`lrm:Activity`) - represents a Digital Ecosystem activity that may be executed during a digital item's lifespan. An activity can be defined as a temporal action that affects, changes, targets or refers to an item. The A&M domain ontologies extend the Activity class, in order to model domain-specific activities (like for example *creation*, *acquisition*, *storage*, *access*, *display*, *copy*, *maintenance*, *loan*, *destruction* of a DO), limiting the list to those that are considered to be important for digital preservation processes in the domain of interest.

**Agent** (`lrm:HumanAgent`, `lrm:SoftwareAgent`) - represents the entity that may perform an activity or may bring change to the digital ecosystem. In the A&M domain, human agents are additionally specialised for the A&M domain into artists, creators, programmers, museum staff etc., and software agents into programs, software libraries, operating systems, etc.

**Dependency** (`lrm:Dependency`) - indicates the association or interaction of two or more resources within the digital ecosystem that may further affect the functioning or display or existence of a DO. In the A&M ontologies, in order to model complex relationships between resources within the context of each subdomain, we extend the basic notion of `lrm:Dependency` into:

- **Hardware dependency**, which specifies the hardware requirements for a resource.
- **Software dependency**, which indicates the dependency of a resource or activity on a specific software agent.
- **Data dependency**, which implies the requirement of some knowledge, data or information (e.g. passwords, configuration files, input from web service, etc.).

Looking in more detail into each of the three ontologies, one may find several design choices regarding notions, properties and restrictions that apply in the context of each subdomain, descriptions of which are given in the following subsections.

**The A&M DVA Ontology**

Digital Video Art is the type of art which contains digital video(s) as its basic concrete part, as well as all the components that a digital video may comprise. A general overview of the main classes and relevant properties of the DVA ontology can be seen in Fig. 3-1. The conceptualization (idea) of a DVA artwork is represented via the `dva:DigitalVideoArt` class (subclassOf `lrm:AbstractResource`) while the actual concrete resource of the digital video is represented via the `dva:DigitalVideo` class (subclassOf `lrm:ConcreteResource`).

Many of the digital video components (i.e. video codecs, video containers, video/audio streams) play significant role in the digital preservation of the DO and involve DP-risks due to massive technological evolution. For that reason, it is important to have in the DVA ontology distinct declarations of relevant classes (`dva:Codec`, `dva:Container`, `dva:Stream` with subclasses of `dva:AudioStream`, `dva:VideoStream`, `dva:SubtitleStream`) and of relevant properties (`dva:hasCodec`, `dva:hasContainer`, `dva:hasStream`, etc.) that connect a digital video with such resources.

There are also significant parameters, considered in DVA as video descriptors (`dva:VideoDescriptors`) that thoroughly define in more detail the exact technological characteristics of digital video files and their components. More specifically, the descriptors which were declared in DVA through relevant classes and properties are the following: *aspect ratio*, *bitrate*, *frame rate*, *chroma format*, *compression type*, *scan type*, *YUV sample range*, etc. The exhaustive list was formalized in our proposed ODP, as mentioned later in paragraph '*Implementation*'.

The actual use and maintenance of DVAs, as stated by CH experts, led us to define several types of activities in the DVA ontology (*creation*, *acquisition*, *storage*, *access*, *display*, *copy*, *maintenance*, *loan* and *destruction*) that may potentially be involved in analysing DP related risks. Activities can be

performed by Agents (via `lrm:performs/performedBy`). They may also be connected with different types of dependencies in order to describe their necessity(ies) on other resources, like, for example, the case where a playback activity requires a specific media player in order to properly display a digital video file of an artwork.

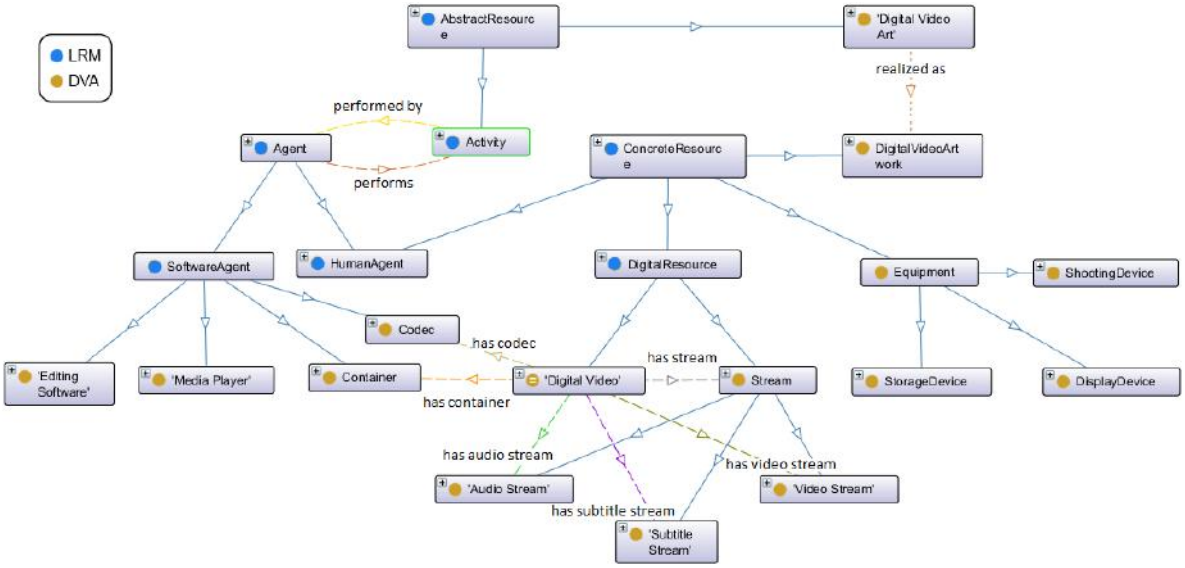


**Fig. 3-1**. Main classes and relevant properties declared in the DVA domain ontology.

Resources used in activities performed, with respect to the artwork, could be part of the `dva:Equipment` class (`dva:DisplayDevice`, `dva:StorageDevice`, `dva:ShootingDevice` etc.) or also of `lrm:SoftwareAgent` (`dva:MediaPlayer`, `dva:EditingSoftware`, etc.).

For class `lrm:Dependency`, the definitions of all three types of specialised dependencies, as well as their connections with other resources, remain the same (as seen in previous subsection). The same stands for the extensions of the `lrm:HumanAgent` class.

### The A&M SBA Ontology

Software-based art is the type of art where the creation of some software plays an important role in the final realisation of the artwork. In the SBA ontology, abstract and concrete parts of the artwork are represented via `sba:SoftwareBasedArt` and `sba:SoftwareBasedArtwork` correspondingly, connected also via the `lrm:realizedAs` property. In order to deal with specialisations of software based artworks [Dekker et al., 2015], we additionally declared four relevant classes, as subclasses of `dba:SoftwareBasedArtwork` (see details in Fig. 3-2).

In the SBA ontology, subclasses of `lrm:Activity` differ from other A&M domains: actions like *compiling*, *emulation*, *conservation*, *migration* and *virtualization*, specialize the content of the SBA ontology to conform with the needs of the domain. As already mentioned in DVA, activities may be connected with agents to represent cases like: "*agent X did the activity Y*", or with dependencies to declare, for example, cases like: "*activity Z depends on resource L in order for the activity to be performed efficiently*".

Since software programs (applications) are the core notion in this ontology, there is a special need to enrich the content of `lrm:SoftwareAgents` and declare relevant resources as its subclasses, to represent the various *compilers*, *APIs*, *software libraries*, *programming tools*, *databases*, etc. that are part of artworks or are involved in activities performed through the lifespan of artworks.

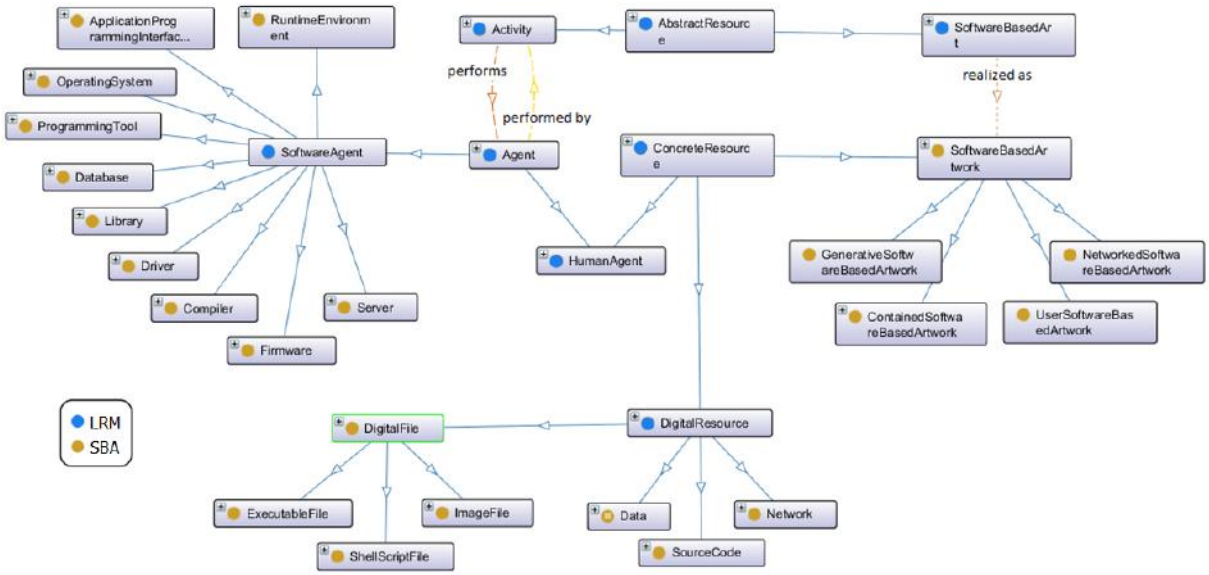


**Fig. 3-2**. Main classes and relevant properties declared in the SBA domain ontology.

For the class `lrm:Dependency`, all three types of specialised dependencies, as well as their connection with other resources, remain the same as declared in the other subdomains. The same stands for the extensions of `lrm:HumanAgent` class.

**The A&M BDA Ontology**

Born-digital archives are personal archives (digital records) from artists, critics or other involved individuals, or relevant company archives from CH institutions and galleries. The main entities in DVA are classes `bda:Fonds`, `bda:Series`, `bda:File` and `bda:Item`, that represent the different levels of description in the domain; this description is based on the General International Standard Archival Description (ISAD(G)) [ICA, 2000], which defines a hierarchical model of the levels of arrangement, with different degrees of detail.

The basic notions in BDA can be seen in Fig. 3-3. In more detail, the born-digital archives are digital resources that may be further classified as *letters*, *emails*, *digital videos*, *photographs*, etc. (of type `lrm:ConcreteResource`), or may be aggregated in *series*, *fonds* or *files* (of type `lrm:AggregatedResource`).

Items and aggregations of items can be accessed, processed, altered and maintained by agents (`lrm:Agent`), either of human or software type, via various types of activities; some BDA-specific activity types are `bda:AppraisalActivity`, `bda:AccessioningActivity`, `bda:CataloguingActivity`, `bda:IngestActivity` and `bda:RedactionActivity`. Special equipment (`bda:Equipment`) may be used for the purposes of each activity, such as computers and several types of storage devices, for which relevant classes exist in BDA ontology.

Again, the definitions of `lrm:Dependency` and its specialisations still remain the same. For the `lrm:HumanAgent` class, additional classes were created to include information for specialised researchers (*art/historian*, *picture* and *provenance researcher*), *archivists* and *conservators*. Concerning class `bda:Group` (subclass of `lrm:HumanAgent`), further specialised classes were created for involved parties (like for example *companies*, *copyright and legal departments* and *funding bodies*).

According to ISAD(G), there are mandatory fields (for example, *ID*, *accession number*, *repository*, *date*, *title*, *description*, *level of description*, *system of arrangement*, *location*, *access status*, *admin history*, *custodian history*, etc.) that need to be defined properly for the archival material instantiated

in BDA, when cataloguing process is applied; these mandatory fields are represented in the BDA ontology by corresponding datatype properties.

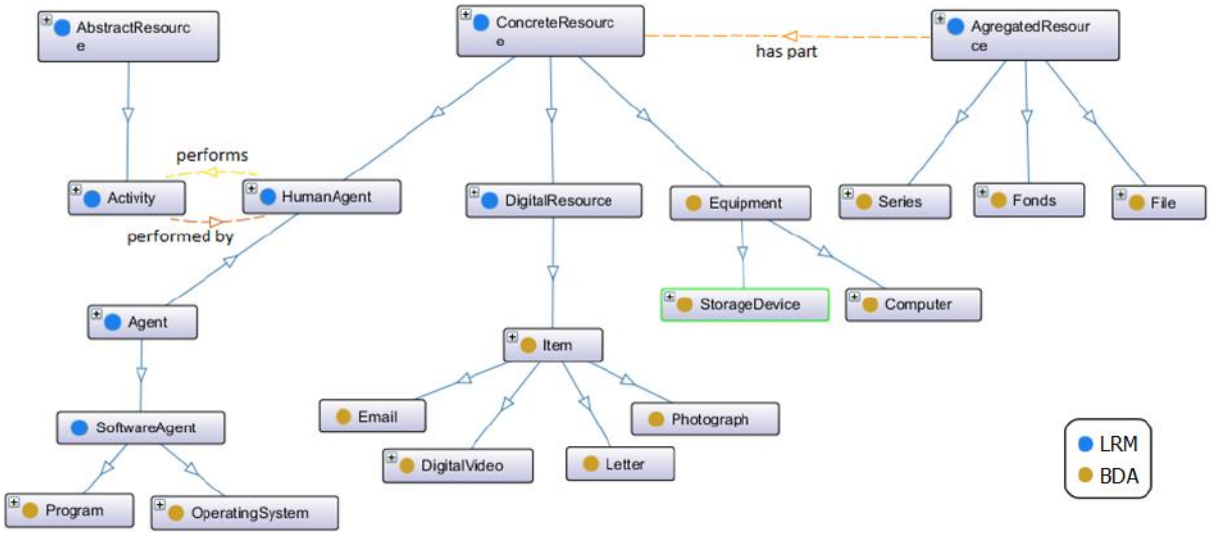


**Fig. 3-3**. Main classes and relevant properties declared in the BDA domain ontology.

**The Space Science Domain Ontology**

Similarly to the A&M domain ontologies, the Space Science domain ontology is also based on LRM constructs to a great extent. LRM dependencies are also specialised in the Space Science ontology as follows (for more details, see also [PERICLES D2.3.2, 2015]):

- **Documentation dependency**: the dependent item is explained by the dependee - when the dependee is not available, the dependent item is not understood.
- **Viewing dependency**: the dependent item is rendered by the dependee - when the dependee is not available, the dependent item cannot be viewed.
- **Processing dependency**: the dependent item is processed by the dependee - when the dependee is not available, the dependent item cannot be processed.
- **Technical dependency**: the dependee is needed for running or compiling the dependent item - when the dependee is not available, the dependent item cannot be run.
- **Non-technical dependency**: the way an environment must be set up to be able to run the dependent item.
- **Configuration dependency**: the dependee is needed as an input to the dependent item - when the dependee is not available, the dependent item cannot be configured.
- **Policy dependency**: a policy dependency defines how two or more items depend on each other, as enforced by a policy, regardless of any technical or other type of dependency.
- **Tacit knowledge dependency**: the dependent item or a process related to it is explained by a person - when the person is not available, the dependent item cannot be understood or used.

The ontologies of the two domains have been developed in parallel, but with occasional communication for aligning the models, despite the vast differences in the application areas. Thus, there are apparent analogies between the dependency types in the two domains, as seen in Table 3-2.

The usage of these types suggests that instead of having a specific association type as a subtype of the “dependency” topic, it can be a subtype of one or more of these dependency types, that in turn will be subtypes of the “dependency” topic. Note that the above list of dependencies may still be refined further within the WP5 and WP6 activities.

**Table 3-2**. Analogy between dependency types in the Space Science and A&M domains.

| Space Science dependencies | Analogous A&M dependency |
|---|---|
| Documentation<br>Configuration<br>Policy<br>Tacit | Data |
| Viewing<br>Processing<br>Technical | Software |
| Non-technical | Hardware |

The rest of the concepts adopted from LRM for semantically representing content include **HumanAgent** and **SoftwareAgent**, which can also be used in the respective Topic Maps hierarchies, having an analogous set of classes in the A&M domain ontologies. Finally, **aggregation** is materialized in the Space Science ontology using "**contained**" associations or "**part-of**" associations. In order to support transitivity, we define these association types as subtypes of the transitive type.

### *FORMALISATION*

For the development of the A&M domain ontologies, we studied the most well-established methodologies in Ontology Engineering (see [PERICLES D2.3.2, 2015]) and then selected the ***NeOn methodology*** [Suárez-Figueroa et al., 2012] as the most suitable and flexible method to follow. The NeOn methodology is a well-structured and exhaustively documented methodology, providing detailed guidance for all key aspects of the ontology engineering process. It can be completely adaptive to application requirements (people involved, end-users, domain(s) of interest). Its main advantage is the ability to cover complex ontology development scenarios, in which the reuse and reengineering of developed knowledge resources and the potential adoption of existing established ontologies and ontology design patterns are promoted.

In order to develop the A&M ontologies according to NeOn standards, we selected to reuse and extend existing CH ontologies (i.e. CIDOC CRM, CRMdig and DC - for more details see paragraphs '*Cultural Heritage Ontologies*' and '*Other Relevant Vocabularies*' in Section 3.1.2) and abstract ontologies (i.e. LRM [PERICLES D3.2, 2014] and DEM [PERICLES D5.2, 2015]) that were developed within the PERICLES project. We created the so called Ontology Requirements Specification Document (ORSD) for each A&M ontology, so as to identify and collect requirements that the ontologies should fulfill, focusing on the following aspects [Suárez-Figueroa et al., 2009]: *purpose*, *scope*, *implementation language*, *intended end-users*, *intended uses*, *non-functional* and *functional ontology requirements*, and *pre-glossary of terms*. The ORSD serves as an agreement between ontology engineers, domain experts and end-users, regarding the explicit requirements that the developed ontologies should satisfy. In order to implement the proposed methodology, a set of relevant documents (case studies, data surveys, competency questions, etc.) were prepared within the PERICLES project, in collaboration with A&M domain and ontology engineering experts.

#### Underlying Formalism

The A&M ontologies are expressed in **OWL**, which is based on **Description Logics** (**DL**s), a family of knowledge representation formalisms characterised by logically grounded semantics and well-defined reasoning services. The main building blocks of DLs are *concepts* representing sets of objects (e.g. `Person`), *roles* representing relationships between objects (e.g. `worksIn`), and *individuals*

representing specific objects (e.g. `Alice`). Starting from *atomic* concepts, such as `Person`, arbitrary complex concepts can be described through a rich set of *constructors* that define the conditions on concept membership. For example, the concept `∃hasFriend.Person` describes those objects that are related through the `hasFriend` role with an object from the concept `Person`; intuitively this corresponds to all those individuals that are friends with at least one person.

A DL knowledge base typically consists of a *TBox T* (**terminological knowledge**) and an *ABox A* (**assertional knowledge**). The TBox contains axioms that capture the possible ways in which objects of a domain can be associated. For example, the TBox axiom $Dog \sqsubseteq Animal$ asserts that all objects that belong to the concept Dog, are members of the concept `Animal` too. The ABox contains axioms that describe the real world entities through concept and role assertions. For example, `Dog(Jack)` and `isLocated(Jack,kitchen)` express that `Jack` is a dog and he is located in the kitchen.

As already mentioned, the implementation language selected for developing the A&M ontologies is **OWL 2** [W3C, 2012]. OWL is a knowledge representation language widely used within the Semantic Web community for creating ontologies. The design of OWL and particularly the formalisation of the semantics and the choice of language constructors have been greatly influenced by DLs. The strength of DL lies in subsumption reasoning and consistency checking and is often applied to classification tasks, i.e. for building or setting taxonomies according to concept and relation definitions [Baader et al., 2003].

Our basic aim is to take advantage of the wide adoption of OWL 2, as well as its formal structure and syntax. There are numerous existing third-party tools/frameworks that support the development of ontologies, such as Protégé[22] and TopBraid[23]. Besides formal semantics, DLs come with a set of powerful reasoning services, for which efficient and complete reasoning algorithms with well understood computational properties are available; for example state-of-the-art implementations include FaCT++ reasoner[24], Hermit reasoner[25], Pellet reasoner[26], etc. The evaluation of ontologies' consistency can be additionally supported by widely-used query languages such as SPARQL[27] or SPIN[28].

**Class and Property Restrictions**

OWL 2 enables the representation of knowledge of a domain by using, apart from entity declarations (classes, properties, individuals), also class expressions and property restrictions. Restrictions should be considered as part of the meaning of a class or property, and thus they participate in the classification process of entities or in the membership definition of individuals in a class.

In the A&M domain, we have defined some common restrictions in entities belonging to all three subdomains, which are described subsequently according to the principles of *Manchester OWL syntax* [Horridge et al., 2006]. Specifically, for the case of `Dependency`:

`HardwareDependency` ≡ `lrm:Dependency` **and** (`lrm:from` **some** `dva:Equipment`)

`SoftwareDependency` ≡ `lrm:Dependency` **and** (`lrm:from` **some** `lrm:SoftwareAgent`)

where, the first declaration reads as "*an instance of class `lrm:Dependency` that has at least one connection to an instance of a* `dva:Equipment` *via property `lrm:from` should be considered as a `HardwareDependency`*", while the second declaration reads as "*an instance of class*

[22] http://protege.stanford.edu/
[23] http://www.topquadrant.com/tools/ide-topbraid-composer-maestro-edition/
[24] http://owl.man.ac.uk/factplusplus/
[25] http://www.hermit-reasoner.com/
[26] https://github.com/Complexible/pellet
[27] https://www.w3.org/TR/rdf-sparql-query/
[28] http://spinrdf.org/

*lrm:Dependency that has at least one connection to an instance of a lrm:SoftwareAgent via property lrm:from should be considered as a SoftwareDependency*". These restrictions are called *existential quantifications* and define a class as the set of all individuals that are connected via a particular property to another individual which is an instance of a certain class. The terms "some" or "one" (the former was used above) are considered the natural language indicators for this type of restriction in the Manchester OWL syntax.

Moreover, for the case of Activity and for its specialised subclasses, we defined a triple of the form <*subject-predicate-object*> that connects an instance of Activity (subject) to an instance of a Resource (object) via a relevant property (predicate). We present some definitions below, while relevant declarations exist for all defined subclasses of the concept Activity:

AccessActivity ≡ accessesResource **some** dva:Resource

CreationActivity ≡ createsResource **exactly 1** dva:Resource

CopyActivity ≡ (hasCopyOutput **some** dva:Resource) **and** (hasCopyInput **exactly 1** lrm:Resource)

A further example of existential quantification from the DVA ontology is the following:

DigitalVideo ≡ hasVideoStream **some** VideoStream

which indicates that a digital video should include at least one digital video stream.

Another property restriction, called *universal quantification*, is used to describe a class of all individuals whose values for a given property belong to a specific class. Universal quantification is declared by the use of terms like "only", "exclusively" or "nothing but", which are the natural language indicators for its usage in Manchester OWL. In the DVA ontology, we have specified that:

DigitalVideo ≡ hasAudioStream **only** AudioStream

DigitalVideo ≡ hasSubtitleStream **only** SubtitleStream

which means that individuals of the class DigitalVideo may or may not have relation to instances of AudioStream and/or SubtitleStream via properties hasAudioStream and hasSubtitleStream correspondingly. In conclusion, only a declaration of the form <instance_1 hasVideoStream video_stream_1> is mandatory in order for the subject of the triple (instance_1) to be considered as DigitalVideo.

Additionally, we have defined property restrictions, in the form of Domain (rdfs:domain) and Range (rdfs:range) declarations. According to OWL 2:

- the rdfs:domain is an instance of rdf:Property that is used to state that any resource that has a given property is an instance of one or more classes, as those are defined in the object of the triple <P rdfs:domain C1>. Anything that is related by *P* to something else, must be a *C1*.
- the rdfs:range is an instance of rdf:Property that is used to state that the values of a property are instances of one or more classes, as those are defined in the "object" of the triple <P rdfs:range C2>. Anything to which something is related by *P* must be a *C2*.

Hence, for every property in the Art & Media domain, specific restrictions in their domain and range declarations were defined in [PERICLES D2.3.2, 2015] (Sections 8.2, 8.4 and 8.6), enriching this way the structure, the semantics and the context of the ontologies.

**Space Science Domain Ontology – Underlying Formalism**

Contrary to the OWL-based A&M ontologies, the Space Science domain ontology is formalized based on Topic Maps (ISO/IEC 13250). The rationale behind choosing Topic Maps instead of OWL as the underlying representation model was based on the fact that the partner responsible (SpaceApps) has

core expertise in this formalization and, also, Topic Maps are extremely user-friendly, making it easy for humans to understand a represented domain.

The Space Science domain ontology is composed of two parts: (a) a set of hierarchical relations between concepts (in Topic Maps terminology, these are called "topic types"), and, (b) a set of representative Topic Maps instances that further clarify the possible relations between various topic types. More thorough descriptions are featured in [PERICLES D2.3.2, 2015]; here only a brief outline is given.

The hierarchical relationships between topic types are of the kind "type-subtype" (or in other words, relationships fitting the statement "subtype is-a-kind-of supertype". Key entities (i.e. topics) and direct subtypes are (most of them have further subtypes that are omitted here):

- **Person**, representing individuals, with the following key subtypes:
  - **Developer** (individuals involved in development activities);
  - **Scientist** (individuals interested in experiments);
  - **Involved in Mission** (individuals involved in the operational mission);
  - **Data Manipulator/User** (people involved with the Space Science data that do not fit in the previous categories).
- **Institutes and Organizations** with the following subtypes:
  - **Space Agency** (a government-controlled space agency);
  - **Control Center** (operating flight hardware);
  - **Industry Academia** (private or academic institute).
- **Software** with the following subtypes:
  - **Operations Software** (software involved in the operational phase of a mission);
  - **Science Software** (software that processes the science data generated by a mission);
  - **Flight Software** (software that is deployed on the payload on board the ISS);
  - **PI Software** (any software used by the Principal Investigator that is not covered by the previous categories).
- **Document** with the following subtypes:
  - **Operations Document** (any document used in or generated by the daily operations of an experiment);
  - **Regulations Document** (any document that contains applicable and relevant rules and regulations that need to be adhered to);
  - **Interface Procedure Document**, further subtyped to OIP (Operation Interface Procedure) and JOIP (Joint Operations Interface Procedure);
  - **Science Document** (any document related to the science behind an experiment);
  - **Training Document** (any document that can be used for training purposes);
  - **Publications** (Journal and Conference Proceedings);
  - **Minutes Of Meeting**.
- **Activities** with the following subtypes:
  - **Data Processing** (all the activities surrounding data processing);
  - **Solar Activity** (activities related to the operations of the SOLAR payload).
- **Hardware** (all the hardware relevant to an experiment and running/operating an experiment) with the following subtypes:
  - **Ground Hardware Component** (hardware used on earth);
  - **Flight Hardware Component** (any hardware that goes into space);
  - **Vehicle** (vehicles to transport hardware and astronauts to and from the ISS and space).
- **Event**: Several types of events can be distinguished. Of particular interest are **Anomalies**, representing off-nominal events, glitches and errors.
- **Period**, representing time spans.
- **Data** with the following subtypes:

- **Telemetry** (data that comes back from the experiment and from ISS);
- **Science Data** (raw and processed scientific data);
- **Measurements** (various sensor measurements of the on board hardware).

The second part of the Space Science domain ontology consists of the instances that are called "**topics**" and can be typed by one or more topic types. Topics can have multiple names and these names can be typed. Topics can also have so called "occurrences", which are pieces of data, relevant to the topic. Finally, topics can be linked to other topics by so called "associations". These associations can also be typed and the role that each topic association can be made explicit.

### *IMPLEMENTATION*

A thorough presentation of the implementation details of the domain ontologies for both domains is given in Chapter 4 of D2.3.2 [PERICLES D2.3.2, 2015]. Here, we give more details on an implemented Ontology Design Pattern (ODP) for representing digital video resources, a first version of which was described in D2.3.2. We are currently in the process of developing additional design patterns for other core aspects of the PERICLES content representation and our aim is to develop a **library of PERICLES semantic constructs** for assisting unfamiliarised end users in populating the semantic models.

Thus, as already seen in paragraph '*Ontology Design Patterns and Mapping to LRM Concepts*' of Section 3.1.2, ODPs constitute reusable solutions to frequently appearing modelling problems [Gangemi, 2005] and are the extension of software patterns for knowledge acquisition in the Semantic Web. The adoption of ODPs in the development process of an ontology increases the standardization level, reinforces the use of best practices and reusable successful solutions, and leads to the wider acceptance of the developed ontology. Patterns are typically published in ODP repositories like http://ontologydesignpatterns.org.

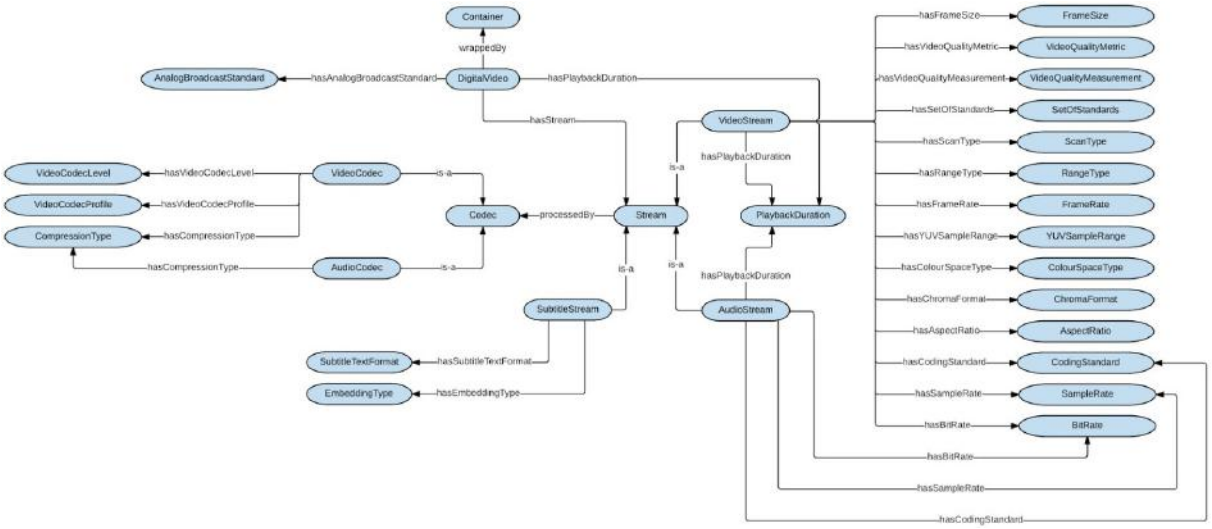


**Fig. 3-4**. Digital Video ODP schematic view.

Within the context of PERICLES we have already developed an ODP, while a few more are also underway. The developed pattern (see Fig. 3-4) is part of the DVA ontology and is aimed at representing digital video resources and their relevant components [Mitzias et al., 2015][29]. The Digital Video ODP intends to model digital video files, their components and other associated entities. Appropriate object properties connect the video file with codecs, streams, containers, etc.

[29] Also available at http://ontologydesignpatterns.org/wiki/Submissions:DigitalVideo

Additionally, the model includes the most significant descriptors for all the previous notions, such as bitrate, aspect ratio, compression type, etc.

The ODP was motivated by the problem of consistent presentation of digital video files in the context of digital preservation and was developed in collaboration with domain experts from Tate. Furthermore, it facilitates the creation of relevant domain ontologies that will be deployed in the fields of media archiving and digital preservation of videos and video artworks. This model was proposed and published at the 6th Workshop on Ontology and Semantic Web Patterns (WOP2015), held at ISWC2015, in order to contribute to the ontology engineering community.

The development of additional ODPs is also underway for other core concepts of PERICLES and will be submitted to relevant research venues (e.g. EKAW).

# 3.2. Context Modelling

## *3.2.1. State-of-the-Art in Ontology-based Context Representation*

The various context modelling approaches found in the literature have been associated with a variety of, often disjoint, application domains. Here, we try to clarify the differences between existing context modelling approaches, an active research field since 2005 [Koç et al., 2014], and to pinpoint similarities, while focusing on approaches that relate to the DP domain.

Existing state-of-the-art surveys have already tried to discriminate context-modelling approaches regardless of application domain. They have identified key modelling approaches that range between *key-value pairs*, *markup*, *graphical*, *object-oriented*, *logic-based* and *ontology-based*, and key requirements, such as *distributed composition*, *partial validation*, *quality of information*, *incompleteness and ambiguity*, *formality and applicability* [Strang & Linnhoff-Popien, 2004]. Another survey concurs with the six aforementioned modelling approaches, but redefines *simplicity*, *flexibility*, *extensibility*, *genericity* and *expressiveness* as requirements [Baldauf et al., 2007]. A more recent study identifies key-value pairs and markup as outdated and less expressive. Thus, it considers modelling approaches as either *object-role based*, *spatial*, *ontology-based* or *hybrid*, while key requirements are *heterogeneity*, *mobility*, *relationships*, *timeliness*, *imperfection*, *reasoning*, *usability* and *efficiency* [Bettini et al., 2010]. A most recent survey gives a statistic measure for the most popular approaches, found to be: *ontology-based*, *graphical* and *logic-based*, followed by *object-oriented* approaches and *markup schemes* [Koç et al., 2014].

Regarding application domains, most existing approaches consider context modelling as a concept tightly linked to context-aware computing, i.e. smart, pervasive environments adapting to user needs. Such systems include pervasive, mobile and ambient intelligence applications, such as smart homes, eHealth, smart office and meeting rooms etc. While the ambient intelligence domain regards real-time adaptive service provision in device network deployments, and is thus seemingly unrelated to DP, some solutions may still be applicable. In detail, DP can take advantage of **modelling individuals, devices and context-of-use** from ambient solutions, disregarding the real-time and service provision aspect. Unfortunately, not many solutions from other domains can be adopted by DP; for instance, the concepts presented in [Mettouris & Papadopoulos, 2013], which regard context-aware recommender systems, are almost entirely disjoint to the concepts required in DP.

Thus, in the framework of the current deliverable, this survey focuses on works that can be related to and applied in the domain of DP, rather than general all-purpose context modelling. Table 3-3 presents a comprehensive comparison of such existing models in literature. The variety of works concurs with the state-of-the-art surveys, as the most dominant modelling approaches are met here, either ontology-based or graphical. Domains of application include pervasive systems, but also museums and eLearning applications.

**Table 3-3**. An overview of existing context modelling approaches.

| Presented in, Year | Modelling Approach | Modelling Language | Application Domain {Concepts} |
|---|---|---|---|
| [Strimpakou et al., 2005] | Graphical | XML | Context-aware services {Person, Service, Preference, Location} |
| [Ou et al., 2006] | Ontology | RDFS/OWL | Context-aware services {Person, Device, Function, Event} |
| [Sheng & Benatallah, 2005] | Graphical | UML | Context-aware services {Location, Language, Temperature, Attraction} |
| [Zhang & Wang, 2005] | Ontology | OWL | Smart Home {Person, Activity, Location, Application, Service, Device} |
| [Gu et al., 2004] | Ontology | OWL | Smart Home {Person, Activity, Location, Application, Service, Device} |
| [Chen et al., 2005] | Ontology | OWL | Smart Meeting {Person, Belief-Desire-Intention, Action, Policy, Time, Space, Event} |
| [Ranganathan et al., 2003] | Ontology | DAML+OIL | Smart Home {User, Device, Service, Location, Time, Weather, Light, Sound, Sports, Health, Mood, Schedule, Activity, Social, Application} |
| [Simons & Wirtz, 2007] | Graphical | UML | Smart Meeting {Person, Activity, Time, Appointment, Meeting, Room} |
| [Chou et al., 2005] | Ontology | RDFS/OWL | Museum {Exhibit, Visitor, Tour Stop, Collection} |
| [Achilleos et al., 2010] | Graphical | UML | Museum {Person, Device, Identity, Time, Location, Activity, Preference, Exhibition, Section} |
| [Van den Bergh & Coninx, 2006] | Graphical | UML | Museum {User, Location, Artwork, Media, Tour, PDA, Screen} |
| [Jovanović et al., 2007] | Ontology | OWL | eLearning {Learning Object, Time, Learning Design, Activity} |

Overall, the most common concepts in context modelling tend to be *Person* and *Device*. Examining approaches per domain, the ones in pervasive computing typically consider *environmental parameters* (e.g. weather, temperature, light and sound), *location*, *user preferences*, *applications* and *services* [Chen et al., 2005; Gu et al., 2004; Ou et al., 2006; Ranganathan et al., 2003; Sheng & Benatallah, 2005; Simons & Wirtz, 2007; Strimpakou et al., 2005; Zhang & Wang, 2005]. Approaches in the museum domain consider smart tour guides, but still model persistent items such as exhibits, exhibitions, artwork and media [Achilleos et al., 2010; Chou et al., 2005; Van den Bergh & Coninx, 2006], as does our proposed model for DP. However, context of use is only captured in [Jovanović et al., 2007], which considers eLearning items used during learning design or certain activities.

## 3.2.2. Modelling of Context in the Art & Media Domain

The most widely referenced definition of context is given by Dey et al. (2001), according to which context is "*any information that can be used to characterize the situation of an entity*", where entity in our domain of interest could be any digital object. For the needs of the DP field and of the A&M domain, we propose in [Kontopoulos et al., 2016] a novel, ontology-based representation approach for modelling context and use-context of digital resources. At the core of the proposed representation lies the LRM.

### SEMANTIC REPRESENTATION OF CONTEXT

We represent context via associations between key classes `lrm:Agent`, `lrm:Activity` and `lrm:Resource`, as shown in Fig. 3-5. More specifically, agents are related to activities via property `lrm:executes` and its inverse property `lrm:executedBy`. Additionally, when relating an activity to a resource, the latter can be either (a) the resource that is affected by the activity and it is indicated by object property `:targetsResource` (inverse of `:targetedByActivity`), or (b) a resource that was used during the activity execution, indicated via object property `lrm:used` (inverse of `lrm:usedBy`). In other words, a targeted resource is the one mainly handled by the activity (e.g. created, borrowed, destroyed), while used resources are those manipulated for the activity execution (e.g. equipment, software, hardware, etc.) and are indicated via object property `lrm:used` (inverse of `lrm:usedBy`).

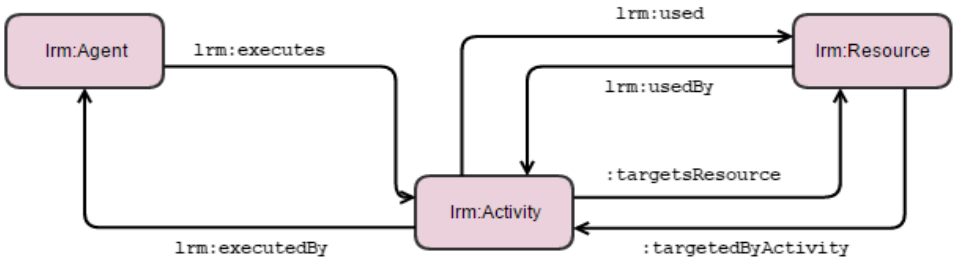


**Fig. 3-5**. Associations between key classes in A&M domain ontologies.

Deltas (`lrm:RDF-Delta`) are another example of LRM notion that can be used to represent context in the domain of interest, and more specifically, to describe changes of resources. These changes may potentially affect other resources as well in the digital ecosystem. The information carried with deltas, may reveal who was responsible for a change in the digital ecosystem, or also how the changes occurred in the system (step-by-step alterations) may affect the system and which was the resulted state of the change in the system. As presented in [PERICLES D3.3, 2015], deltas give meta-information about the modification of a resource, by defining a list of triples that have been deleted as well as a list of triples that have been inserted (through the use of `rdf:Statement` meta-descriptor and of properties `lrm:deletion` and `lrm:insertion` correspondingly). It should be noted that an `rdf:Statement` is the statement made by a token of an RDF triple; the subject and the object of the `rdf:Statement` are instances of `rdfs:Resource` (and `lrm:Resource`) that are identified through the use of `rdf:subject` and `rdf:object` in the triple correspondingly, while the predicate of the `rdf:Statement` is an instance of `rdf:Property` that is identified through the use of `rdf:predicate` in the triple. A simple representation scheme of `lrm:RDF-Delta` and its involved notions and properties can be seen in Fig. 3-6.

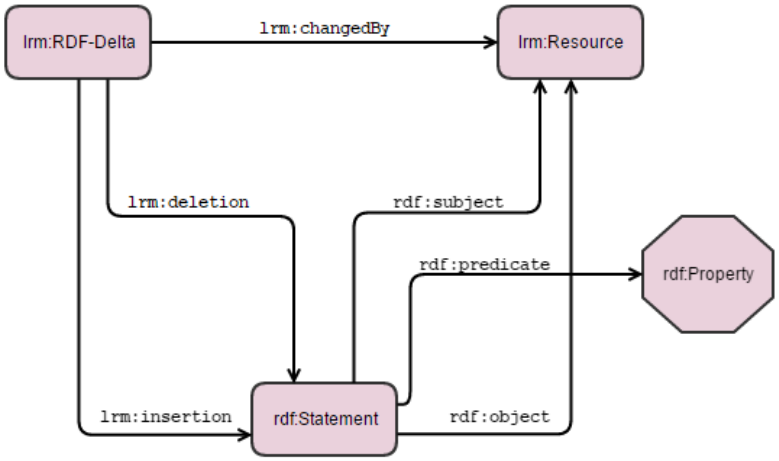


**Fig. 3-6**. Representation of `lrm:RDF-Delta` class in LRM.

### *Semantic Representation of Use-Context*

Regarding the representation of use-context, we utilize the `lrm:Dependency` which is explicitly augmented with rich semantics, for modelling the underlying preconditions, intentions, specifications and impacts. Thus, dependencies constitute meaningful correlation links among resources and use-contexts. The notion of *intention* specifies what a dependency intends to express and *specification* thoroughly describes the dependency itself and its context. Furthermore, the notion of *precondition* describes the contextual properties that need to hold in order to consider the dependency as "activated", and the notion of *impact* describes what actions follow when the dependency is activated [PERICLES D3.3, 2015; Lagos & Vion-Dury, 2016].

In order to turn dependencies into meaningful correlation links among resources and use-contexts, we have added a set of predefined intention types in order to represent all relevant dependency occasions seamlessly. Below is a description of the proposed intention types [PERICLES D4.3, 2016]:

- **Dependencies with a conceptual intention** are aimed at modelling the intended "meaning" of a resource (i.e. an artwork), according to the way the creator meant for it to be interpreted/understood. For example, a poem (digital item) belonging to an archival record may not conserve its formatting during the normalization process, something that is against the intention of the poet regarding the way that the poem is conceptualized/conceived by a reader.
- **Dependencies with a functional intention** represent relations relevant to the consistent and complete operation/functioning of the resource. For example, a specific codec is required to display a digital video artwork.
- **Dependencies with a compatibility intention** model components which may operate together or as replacement for obsolescence, lack of availability or other reasons. For example, the software used for playing back a digital video artwork is compatible with certain operating systems.

### *Implementations*

An instance of `lrm:Dependency` may represent the necessity of existence/use of specific resource(s) in order for a DO or other resource to operate efficiently. Instantiations of dependencies can be seamlessly implemented for the real case scenarios [PERICLES D2.3.1, 2014; PERICLES D2.3.2, 2015] presented by experts in the DP and CH domain; these implementations may stand as single instantiations or may be part of a chain of dependencies, with further conjunctive or disjunctive roles; a *conjunctive dependency* requires all dependent entities to be present, whereas a *disjunctive*

*dependency* requires at least one of a set of entities to be present [PERICLES D3.2, 2014; PERICLES D3.3, 2015]. A demonstration of chaining dependencies for a real case scenario, using also conjunctive and disjunctive relationships, is presented in Fig. 3-7.

More specifically, an instance of digital video (`digital_video_1`) is connected with a specific type of container (`AVI`) via the property `dva:hasContainer`. The instance of `AVI` is part of a software dependency, which declares that the specific container can be manipulated properly via three different media players. This dependency is of `:SoftwareDependency` type and the context of use is specified via the `compatibility` intention; this instance of dependency is additionally of `lrm:DisjunctiveDependency` type, which means that (at least) one of its resources declared in `lrm:from` property is necessary in order for the declared resource in `lrm:to` to function properly, as the intention and specification defines.

As we continue with the chain representation, we can mention that there is a specific media player (that is `Windows Media Player` in our ontology) that is compatible with specific versions of Windows Operating System; again, this disjunctive relation is represented as a `:SoftwareDependency` (see `software_dependency_3` in Fig. 3-7) with `compatibility` intention.

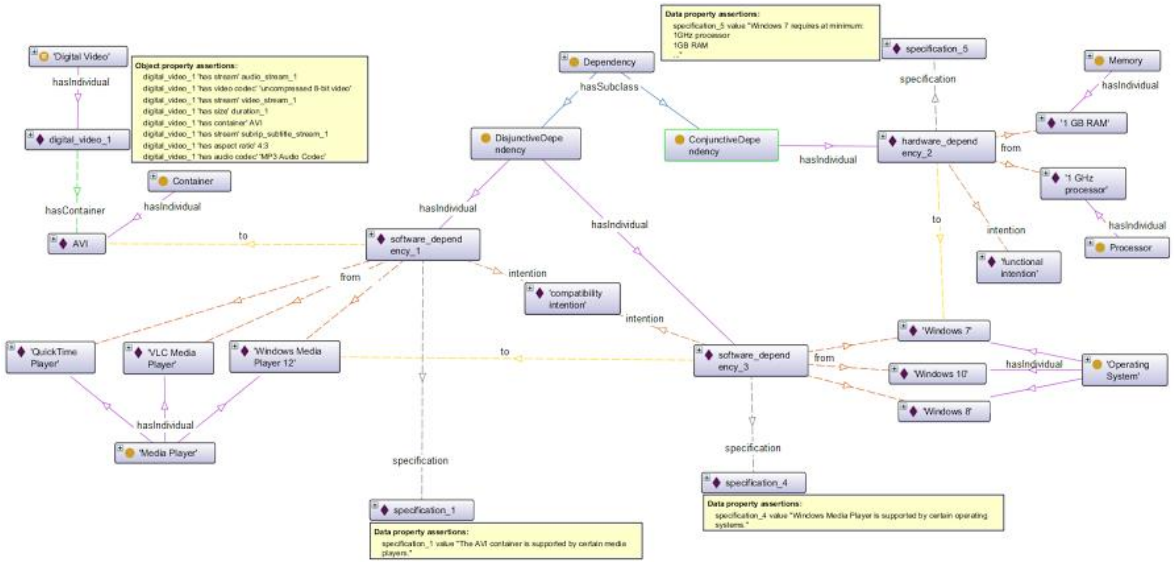


**Fig. 3-7**. Representing a chain of dependencies in DVA.

Finally, we conclude in this example, with a `:HardwareDependency` (see `hardware_dependency_2` in Fig. 3-7) that presents the dependency of a specific version of Windows OS (`Windows 7`) on specific instantiations of hardware equipment (i.e. memory and processor), supposing that these are the minimum requirements in order to install/run Windows 7 efficiently. The context of use under which this specific dependency is defined, is `functional`, and such is the type of intention. This instance of dependency is additionally of `lrm:ConjunctiveDependency` type, which means that all of the resources declared in `lrm:from` property are necessary in order for the declared resource in `lrm:to` to function properly, as the intention and specification defines.

Regarding deltas, we may consider an example involving the change of a media player that is used for playing a digital video file; in our ontology (i.e. DVA), we model the state that "a video playback activity uses a specific media player" in order to reproduce some digital video, via corresponding classes as seen in Fig. 3-8. In order to represent the process of selecting another player for the video file, we substitute the media player in the `lrm:used` relationship of the particular activity.

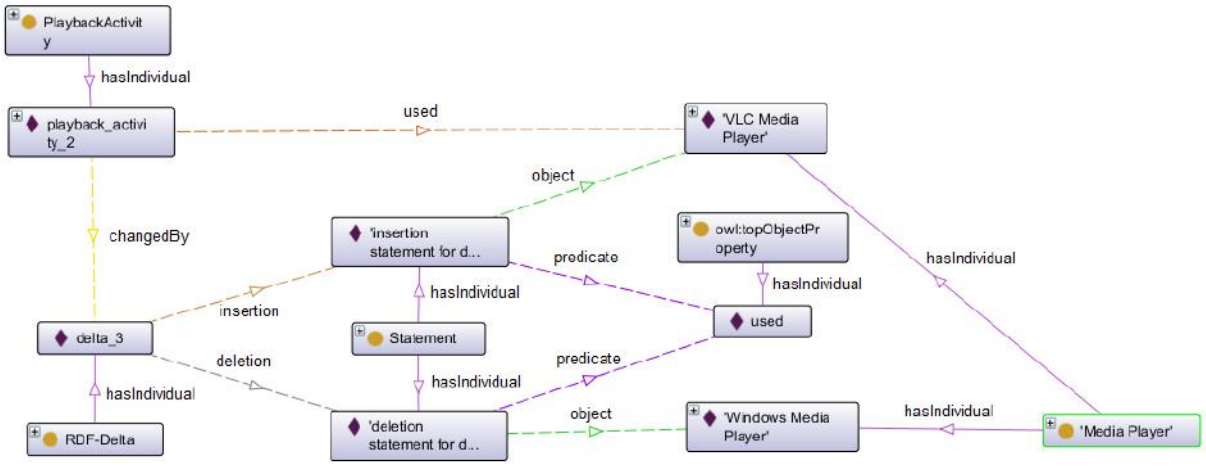


**Fig. 3-8**. A media player change in playback activity represented with a simple delta.

This change is represented with an `lrm:RDF-Delta` instance (`delta_3`) connected with `playback_activity_2` through property `lrm:changedBy` (see Fig. 3-8). Additionally, `delta_3` incorporates two statements (`rdf:Statement`) via properties `lrm:deletion` and `lrm:insertion`, indicating that the statement `<playback_activity_2 lrm:used 'Windows Media Player'>` was replaced in the ontology by the statement `<playback_activity_2 lrm:used 'VLC Media Player'>`. The final state of the model is that the specific playback activity is now performed with the use of VLC Media Player.

Note that the deltas examples presented above have been deployed in developing the first version of the **MICE** tool (**Model Impact Change Explorer**), which is a core output of WP6.

## 3.3. Contextualised Content Semantics

As mentioned before in this document, the term "contextualised semantics" refers to variations in meaning and interpretation that arise according to the context in which content is viewed. This subsection focuses on our adopted methodologies for taking advantage of context representation towards creating an additional "layer" of inference on-top of the developed models. More specifically, the described ontology-based OWL representations allow automatic reasoning and handling of various inconsistent cases that are significant within the context of the domain of interest. Here, we present an implementation of such a validation layer, which is based on the DVA ontology and uses the **SPARQL Inferencing Notation** (**SPIN**) [Knublauch et al., 2011]. A more thorough account of this line of research will be given in the upcoming deliverable D4.5 (due M44).

SPIN is a well-known notation for representing SPARQL rules and constraints on models, for performing queries on RDF graphs. SPARQL queries can be stored as RDF triples alongside the RDF domain model, enabling the linkage of RDF resources with the associated SPARQL queries, as well as their consequent sharing and reuse. SPIN can also be used to derive new RDF statements from existing ones through iterative rule application.

In the A&M ontologies, SPIN rules are used for taking advantage of elements from the context of digital resources in order to detect inconsistencies while examining a specific state of the digital ecosystem, or for cases where SPIN rules monitor policies existing in the digital ecosystem in order to trigger changes that policies describe. Examples of both cases are given in the following subsections: two representative evaluation scenarios taken from [Rice, 2015], that have been implemented through relevant SPIN rules [Lagos et al., 2016], and also one advanced example that expresses precondition and impact of dependencies as SPIN rules, that track a policy of a real case scenario and perform a change in the ecosystem accordingly.

It should be noted that, for the case of inconsistency checking (use cases 1 and 2), we classify in the ontology the inconsistent or 'problematic' instances as error (`dva:ErrorItem`[30]) or warning (`dva:WarningItem`[31]) entities, incorporating at the same time corresponding descriptive message fields (i.e. properties `dva:hasErrorText` and `dva:hasWarningText`) that specify the nature of the problem.

**Use Case 1 - Detect Inconsistency in Container's Metadata (no Aspect Ratio Information).**

This scenario aims to detect whether a container's metadata carry the aspect ratio information (i.e. 4:3, 16:9, 21:9) of a digital video, which is necessary for the consistent playback of video files. It is possible that some types of containers do not include information on the aspect ratio value of the digital video, even though this information may already be known by the (human) creators or owners of the files.

As an example, we consider two digital video files wrapped by different container types. In the DVA ontology, this information is represented with the following triples:

```
?digital_video_1    a                   dva:DigitalVideo
?digital_video_1    dva:hasContainer    ?avi
?avi                a                   dva:Container
?avi                rdfs:label          'AVI'
?digital_video_2    a                   dva:DigitalVideo
?digital_video_2    dva:hasContainer    ?matroska
?matroska           a                   dva:Container
?matroska           rdfs:label          'MATROSKA'
```

where `dva:DigitalVideo` is a subclass of `lrm:DigitalResource` and `lrm:ConcreteResource`, and `dva:Container` is a subclass of `lrm:SoftwareAgent`.

Due to limitations of the AVI container, the aspect ratio of `digital_video_1` is not stored in the file's metadata; this information can be represented in the ontology with the following triples:

```
?avi                dva:includesAspectRatio    false
?matroska           dva:includesAspectRatio    true
```

When a playback activity is performed (and captured through corresponding ontology instantiations), the ontology should infer any inconsistency related to unspecified aspect ratio of a digital video in its container's metadata. Thus, in the aforementioned case, `digital_video_1` that has an AVI container will be classified as `dva:WarningItem`. An explanatory text will be attached to the item, via the use of the property `dva:hasWarningText`. The SPIN rule that detects missing aspect ratio values in the digital video's container metadata and classifies relevant resources as warning items, is given below:

```
CONSTRUCT
{
?digital_video      a                       dva:WarningItem.
?digital_video      dva:hasWarningText      "No aspect ratio information in
                                            container".
```

[30] An error item indicates an inconsistency, whose impact may completely affect or prevent the operation/functionality of a resource.

[31] A warning item indicates an inconsistency, whose impact may affect the conceptual/visual output of an action, but not the actual operation/functionality of a resource.

```
}
WHERE
{
?digital_video      dva:hasContainer          ?container.
?digital_video      a                         dva:DigitalVideo.
?container          a                         dva:Container.
?container          dva:includesAspectRatio   false.
}
```

It should be noted that the faulty entity (here `digital_video_1`) is classified as a `dva:WarningItem` and not as a `dva:ErrorItem`, since the digital video will be played but, possibly, not with the proper size/resolution; the media player cannot track the actual aspect ratio of the digital video and it will apply a default value instead.

**Use Case 2 - Detect Inconsistency in Playback Activity (Incompatible Player).**

In this case, SPIN rules check if the available media players of a given system (installation) are qualified to play the available digital video files properly. They demonstrate, in practice, how compatible media players could be detected for certain video files, based on the supported containers defined for each player.

We again consider the aforementioned instantiations of `digital_video_1` and `digital_video_2`, as sample digital video files, and their related containers. Based on the ontology representation below, these containers may be connected with compatible media players through instantiations of class `dva:SoftwareDependency`:

```
?software_dependency_1    lrm:from     ?avi
?software_dependency_1    lrm:to       ?windows_media_layer
?software_dependency_1    lrm:to       ?quicktime_player
?software_dependency_1    lrm:to       ?vlc_player
?software_dependency_2    lrm:from     ?matroska
?software_dependency_2    lrm:to       ?vlc_player
?windows_media_player     a            dva:MediaPlayer
?windows_media_player     rdfs:label   'Windows Media Player'
?quicktime_player         a            dva:MediaPlayer
?quicktime_player         rdfs:label   'QuickTime Player'
?vlc_player               a            dva:MediaPlayer
?vlc_player               rdfs:label   'VLC Media Player'
```

where `dva:MediaPlayer` is a subclass of `lrm:SoftwareAgent`.

By interpreting the above representation manually, we may conclude that `digital_video_1` could be efficiently displayed with any of those three media players, while `digital_video_2` could be played efficiently only with VLC. In order for the ontology to automatically infer a media player incompatibility for a digital video, the corresponding instantiation of a `dva:PlaybackActivity` should be considered, i.e. for example the playback activity instance for `digital_video_2` as seen below:

```
?playback_activity_2      a                 dva:PlaybackActivity
?playback_activity_2      dva:playsResource ?digital_video_2
?playback_activity_2      lrm:uses          ?windows_media_player
```

If the used media player is not defined as compatible with the video's container, then this specific instance of playback activity should be classified as `dva:ErrorItem`. By evaluating the above ontology instantiations, we expect that `playback_activity_2` will be classified as an error item, because it uses `Windows Media Player`, which is not compatible with the container (MATROSKA) of `digital_video_2`. The SPIN rule that checks the compatibility of media players for available instances of playback activity can be seen below:

```
CONSTRUCT
{
?activity    a                          dva:ErrorItem .
?activity    dva:hasErrorText           "Incompatible player for playback
                                        activity".
}
WHERE
{
?digital_video      dva:hasContainer    ?container.
?digital_video      a                   dva:DigitalVideo.
?dependency         lrm:from            ?container.
?dependency         a                   dva:PlayerDependency.
?activity           dva:playsResource   ?digital_video.
?activity           lrm:used            ?player.
MINUS
  {
       ?dependency   lrm:to              ?player.
  } .
}
```

**Use Case 3 - Detecting Violation of a Policy and Perform Change.**

This scenario involves an instance of a digital image (`image_1`) associated with a specific logo (`logo_1`) via a dependency (Fig. 3-9). The background of the scenario is a policy which states that an organisation logo needs to be embedded on all images made available online, and the policy is implemented by a dependency and related precondition-impact. Here, we consider a case where `logo_1` is replaced by `logo_2` (i.e. the image file is left unmodified and, instead, a completely new entity (image+logo) is created and pointed to[32]), according to a respective policy. An instance of delta is created to describe the aforementioned change in the dependency.

The dependency contains further information in *precondition* and *impact*, defining the behaviour of the dependency against change, in the form of SPIN rules. Here, the desired behaviour is to update the image whenever a change in the logo occurs. Thus, *precondition* and *impact* define in essence a policy that says "*every time the logo of an image changes, initiate the impact part*". With the use of SPIN rules and SPARQL query language, the following directions are given:

- the precondition states that "*if the* `lrm:from` *part of the dependency is changed through an instance of delta, then trigger what is described in the impact*", and
- the impact reformulates the dependency, placing the instance of the new image at the `lrm:to` part, and, furthermore, creates a new delta that expresses this change.

[32] An alternative use case would be updating the image file - the logo would change and the file would be updated, maintaining though the same name and file ID. This would result in having the respective LRM entity updated with the new metadata descriptions, making use of LRM's versioning mechanism.

It should be noted that the new delta instance follows the previous one via property `lrm:follows`.

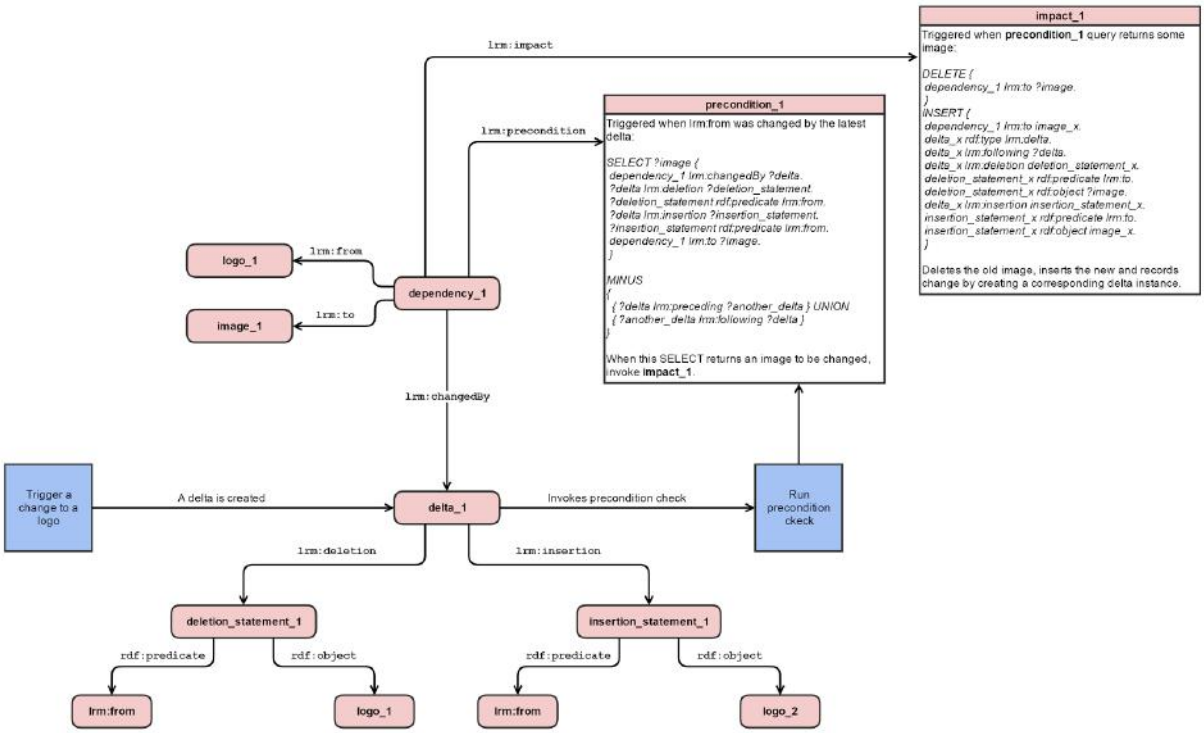


**Fig. 3-9**. A workflow representing a policy to handle change in a dependency.

As seen in Fig. 3-9, precondition and impact may be expressed in SPARQL. Thus, if the SELECT part of `precondition_1` returns some result, then the UPDATE query contained in `impact_1` should be executed. Cumulatively, a single SPIN rule can be generated to combine both functionalities. Such a rule could be added to a list of SPIN rule checks that need to be performed whenever a delta is created (i.e. whenever a change in the digital ecosystem takes place).

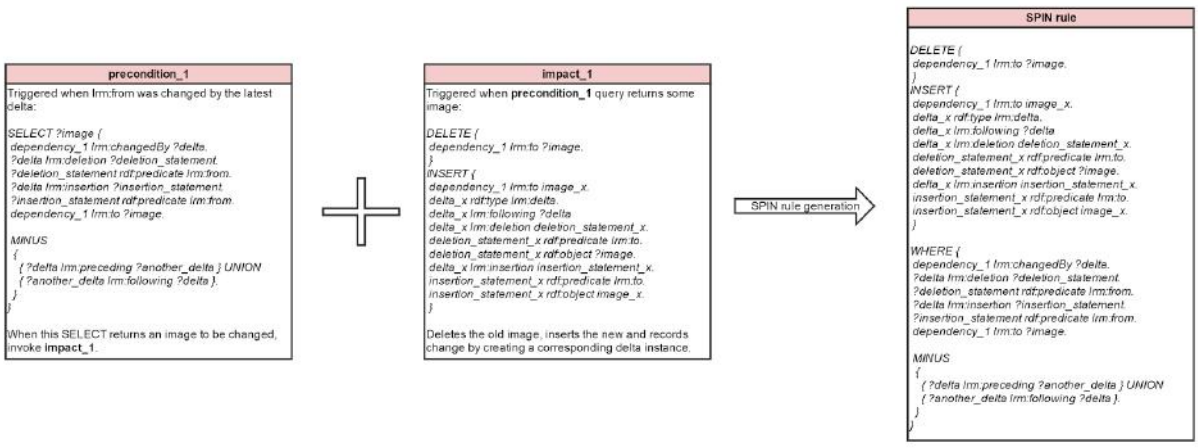


**Fig. 3-10**. Precondition and impact combined into a SPIN rule.

## 3.4. Chapter Summary

This chapter presented our proposed ontology-based approaches for semantically representing content pertinent to the Space Science and Art & Media domains of the PERICLES project. As already mentioned, although the two domains are vastly different, the two ontologies share several common

characteristics, like e.g. dependency subtypes and various categories of agents (human and/or software) and activities. The chapter also presented the adopted methodologies for developing the models, along with the underlying formalisations, which are, again, different (OWL and Topic Maps) but share several commonalities. Additionally, we discussed our proposed approaches for semantically representing contextual information based on the presented ontologies, also providing details on the respective implementations in OWL, by using the DVA domain ontology as the test-bed. Finally, the chapter introduced our proposed methodologies on contextualised semantics, i.e. taking advantage of context representation towards creating an additional inference layer on-top of the developed models. The developed layer presented here is based on the SPARQL Inferencing Notation (SPIN), an established notation for representing SPARQL rules and constraints on ontological models. Indicative examples from the A&M domain have been presented.

# 4. Statistical Context Modelling and Contextualised Content Semantics

After having introduced the notion of contextualised content semantics in the intellectual framework of ontology construction and development in the previous chapter, now we inspect a different approach toward the same goal, ontology maintenance. To distinguish between them, ontology building relies on the power of the human mind and uses logic as its major tool. On the other hand, multivariate statistics provides researchers with computational methods suitable for the automation of this conceptualization process (and several related efforts as the history of information retrieval and machine learning tells us). Ultimately both address the problem of deducing one generally valid set of rules from plenty of examples, with research into automatic ontology extraction connecting their respective contributions (see e.g. [Ren, 2014]). However, in the case of statistics, we use as a basis of analysis datasets with tens of thousands to millions of DOs.

## 4.1. Application of Context and Contextuality in Theories of Meaning for Semantic Spaces

In 21st century, after fifty years of research progress into creating semantic spaces for information retrieval and machine learning, the two most popular mathematical objects used by the statistical approach are **probabilities** and **vectors**. Because the notion of contextuality, detailed below and of importance in upcoming D4.5, is a probabilistic one, we start with an eye on probabilities here and will gradually move over to vectors as our own means of information representation in PERICLES experiments. However, we also note in passing that the two methodologies are conceptually related, since they both adopt a generic interpretation of context, i.e. any information in the environment of the digital resource that tells us something about the situation the resource is in or about its use by and interplay with other resources. Examples of this could be the neighboring words of a term in a piece of text, or the interrelated constructs of a resource in an ontology.

Regardless of its global success, the inherent problem of using statistics for the processing of meaningful entities in general is that one has to express quality by quantity. It's no wonder then that the interplay between probabilities, context, meaning and evolution is rich, subtle and controversial. For a start, there are intellectual problems with Bayesian probability used as a point of departure for the models which end up in modelling context for different purposes. The biggest problem is its success in information retrieval (IR) and related fields. Namely, for an IR model to be successful, its relationship with at least one major theory of word meaning has to be demonstrated; with no such connection, meaning in numbers becomes the puzzle of the ghost in the machine. For the vector space IR model (VSM) – underlying many of today's competitive IR products and services – such a connection can be demonstrated; for others like PageRank[33], the link between graph theory and linear algebra leads to the same interpretation. Namely, in both cases, the theory of word semantics cross-pollinating numbers with meaning is of a contextual kind, formalized by the distributional hypothesis [Harris, 1968]. As a result, the respective models can imitate the field-like continuity of conceptual content, likely to account for their market penetration. However, unless we consider the VSM roots of both the probabilistic relevance model[34] and its spinoffs including BM25[35], such a link is still waiting to be shown between probability and semantics[36].

[33] https://en.wikipedia.org/wiki/PageRank
[34] Because it departs from a "binary index descriptions of documents", see [Robertson & Spärck Jones, 1976].

As we are interested in events embedded in evolving contexts, this constrains the researcher to look at (a) *context and contextuality* as related to Hilbert space also including Euclidean space (with a possible, but not necessarily inevitable detour into kinds (interpretations) of probability); (b) *models of temporality* including *types of change*; and (c) *their respective fit with interpretability*, i.e. semantics, regardless of the actual kind of digital objects. Then the research question is, "*if context leads to contextuality, a property of quantum-like systems, how far can the analytical method be applied to our systems, i.e. how far are they quantum-like?"*. We note in passing that this question is the link between the current deliverable and the upcoming D4.5.

**(a) Context and contextuality**

Based on 3.5.1, we define context as "*any information that can be used to characterize the situation of an entity*" [Dey et al, 2001], in line with the PERICLES glossary definition: "*Context of a digital object is anything external to the object itself that can affect its interpretation*." In a wider sense, we can think of "*The surroundings, circumstances, environment, background or settings that determine, specify, or clarify the meaning of an event or other occurrence*"[37] so that regionality, neighbourhood, thresholding, or window size in text analysis are relevant related concepts as well. Once temporality is added, the notion of context can include causality, conditionality, predecessors and successors etc. As in this and the next chapter we will focus on text and images indexed by natural language, we can further specify semantic context as the surrounding text in which a word or passage appears and which helps ascertain its meaning. Here, semantic context and use context overlap because language use constrains the index term frequency statistics underlying our results. Further, for upcoming D4.5, in Hilbert space, context returns as contextuality by virtue of contextual probability. The idea goes back to Khrennikov [Khrennikov, 2010] and uses the fact that events with contextual probabilities project on subspaces they cannot leave afterwards, i.e. which become their context[38]. Thereby, it is an interpretation option to consider contextual probability as capturing situations, where both successive and separated co-occurrences can record elements of the situation[39]. This is the so-called *propensity interpretation*.

Because of a variety of contexts, all potentially pertinent for information representation (linguistic, social, cognitive, workflow, etc.) [Guha & McCarthy, 2003], context and contextuality are key to improving models of semantic spaces [Widdows, 2004; Baroni et al., 2007] and Digital Preservation alike [Dallas, 2007; Moore, 2008], going back ultimately to the importance of social embedding [Couch, 1992].

**(b) Temporality**

Given that the scientific approach recognizes three major concepts of time, i.e. (1) time as it is to conscious awareness (irreversible/subjective); (2) time as it is to theoretical physics (with no intrinsic direction, i.e. reversible/objective); and (3) time as it is to thermodynamics and the evolutionary sciences such as biology (irreversible/objective) [Denbigh, 1982] strictly speaking it is a matter of

---

[35] See p. 339 in [Robertson & Zaragoza, 2009]

[36] Another proof of correspondence between geometry and probability in Hilbert space is the angular separation of two vectors projected onto a subspace which corresponds to probability: "In the quantum formalism, any event is defined by a subspace. Hence, in the user's information need (IN) space, for each document *d*, we can define a subspace $O_d$ corresponding to the event 'the document *d* is relevant'. If we let |*phi*> be the user's IN state, the probability Pr(R|*d*, *phi*) of the document *d* being relevant to the user's IN |*phi*> is defined as the square of the length of the projection of the vector |*phi*> onto the subspace $O_d$, which adheres to the definition of probabilities in quantum mechanics." [Frommholz et al., 2010].

[37] https://en.wiktionary.org/wiki/context

[38] Classical conditional probabilities do not have such properties, i.e. this method is applicable both to spatial and temporal context modelling.

[39] For various interpretations of probability, see [Hájek, 2012].

study to decide which of the above temporality concepts – if not a blend of theirs – would suit best digital preservation [Raubal, 2008; Bennett & Galton, 2004; Kauppinen et al., 2008].

The notion of time used here and in Section 5.3.1 does not differ from the LRM approach[40] but could be potentially enriched by the above observations[41]. These were listed for the sake of D4.5 which, by looking at physics as a metaphor to model evolving semantics, must include different understandings of the concept, including a backward flow of time as in Feynman diagrams.

**(c) Interpretability**

Finally, having selected a preferred model of temporality, one has to ask if the results can be interpreted. This presupposes a good fit with some reasonably formalized theory of semantics. Here, another two questions emerge: *can the observed features be regarded as entries in a vocabulary*? If so, in this case distributional semantics applies, and, given more complex representations, other types may do so as well [Wittek et al., 2013]. The second question is, *do they form sentences, or does the concept of linear content transmission apply to the digital object, or combinations thereof*? For example, one could regard a workflow (process) a sentence, in which case compositional semantics applies [Coecke et al., 2010; Sadrzadeh & Grefenstette, 2011].

Turning now to theories of *word* vs. *sentence semantics* as studied in general linguistics and semiotics, the first problem is that, due to increasing interest from late 19th century onward but also dating back to Aristotle and even before, there are a great number of theories of both word and sentence meaning. E.g. for word meaning, the starting point for PERICLES, currently one has to consider 5-10 different such theories whose match with geometry or probabilities is only partly resolved. Prominently, the best known theories fall in three major groups:

- "Meaning is use": Wittgenstein's idea [Wittgenstein, 1963] about habitual usage provides indirect contextual interpretation of any term (cf. Harris´ distributional hypothesis [Harris, 1968], see also Firth[42]), connecting this paradigm to de Saussure's structuralism. The frequency of word use, underlying vector space models of information representation, expresses aspects of a conceptual hierarchy, in accord with findings regarding the inverse relationship between the number of features [intensions] an object has vs. the number of objects in its respective class [extensions] by Carnap[43], plus connecting intensional logic with the distributional observations of Zipf's law[44] and its applications for automatic indexing by [Luhn, 1960]); their interplay, expressed in metric space, turns sense relations into a measurable form.
- "Meaning is change": the stimulus-response theory of meaning proposed e.g. by Bloomfield[45] in anthropological linguistics and Morris[46] in behavioural semiotics, plus the biological theory of meaning of Uexküll[47] [Uexküll & Kriszat, 1956] both stress that the meaning of any action is its consequences.

[40] In D3.3 [PERICLES D3.3, 2015] we proposed an LRM-based time construct that encompasses three different temporal concepts: instants, time intervals and durations.
[41] For treating time, the dynamic version of LRM builds on Allen's interval algebra [Allen, 1983]. This calculus specifies 13 base relations that capture the possible relations between two intervals X and Y, by which given facts can be formalized and then used for automatic reasoning. The LRM defines a time instant as a coordinate in time space and embeds mechanisms to handle uncertainty while using the standard astronomical time arrow, i.e. forward progress.
[42] "You shall know a word by the company it keeps" [Firth, 1957:11].
[43] http://plato.stanford.edu/entries/logic-intensional/
[44] https://en.wikipedia.org/wiki/Zipf's_law
[45] https://en.wikipedia.org/wiki/Leonard_Bloomfield
[46] https://en.wikipedia.org/wiki/Charles_W._Morris
[47] https://de.wikipedia.org/wiki/Jakob_Johann_von_Uexk%C3%BCll

- "Meaning is equivalence": direct reference theory[48], Peirce's sign relation[49], or denotational semantics[50] suggest that 'X = Y for/as long as Z'. This holds for any ontology as well.

A particular approach to word semantics is Trier's theory of semantic (or lexical) fields [Trier, 1934]. This is a good candidate for a unification effort because: (1) its 2-dimensional representation of vocabulary units with related meaning complies with the 2-dimensional projection from a higher-dimensional latent semantic space, (2) such a projection goes back to the distributional hypothesis, i.e. is context-dependent; and (3) it is in general suitable to express topical compositions as regions in the plane, regardless whether they go back to single concepts, or bags-of-concepts (i.e. documents), or sequences of concepts (e.g. phrases or sentences). This flexibility is useful to bridge the gap between computational linguistics and vector-, graph- vs. probability-based encodings of semantic content, which, in turn, results in a generic tool for digital preservation.

Briefly, we also have to mention two new kinds of semantics pertinent for the study of evolving semantics, called "**update semantics**" [Veltman, 1996] and "**dynamic semantics**"[51]. By computational linguistics and language philosophy, both have been construed to address, as their names suggest, new sentence semantic problems which are relevant to our current frame of thought. Finally, we have to stress that by moving over from static to evolving semantic fields and represent the latter by a vector field, the original constraint of context-dependence remains.

We believe that the temporal evolution of contexts is a natural way to make progress in modelling. We already demonstrated how animated visualisation of evolving text corpora could display the underlying dynamics of semantic content. To interpret the results, one needs a dynamic theory of word meaning [Daránvi & Wittek, 2013a]. We also suggested that conceptual dynamics as the interaction between kinds of intellectual, emotional etc. content, and language, is key for such a theory, and demonstrated our methodology by two-way seriation – a popular technique to analyse groups of similar instances and their features, as well as the connections between the groups themselves [Daránvi & Wittek, 2013b]. The two-way seriated data were visualised as a two-dimensional heat map or a three-dimensional landscape where colour codes or height corresponded to the values in the matrix. To achieve a meaningful visualisation we introduced a compactly supported convolution kernel similar to filter kernels used in image reconstruction and geostatistics. This filter populated the high-dimensional sparse space with values that interpolated nearby elements, and provided insight into the clustering structure. We also extended two-way seriation to deal with online updates of both the row and column spaces, and, combined with the convolution kernel, demonstrated a three-dimensional visualisation of dynamics. Therefore with the above caveats, we are now ready to take the next steps, first regarding computability (below, in Section 4.2), then for an experiment (in Section 5.3.1).

## 4.2. A High Performance Computing Model of Evolving Semantic Content

As briefly outlined in D4.3 ([PERICLES D4.3, 2016], Section 4.1.3), for a practical analysis of context-dependent correlations leading to automatic classification, we have been developing a high-performance qualitative machine learning algorithm called **Somoclu** ("**S**elf-**O**rganizing **M**aps **O**ver a **Clu**ster") [Wittek et. al, 2015a]. This tool is primarily meant for training extremely large emergent self-organizing maps on supercomputers, but it is also the fastest implementation running on a single

[48] https://en.wikipedia.org/wiki/Direct_reference_theory
[49] https://en.wikipedia.org/wiki/Charles_Sanders_Peirce
[50] https://en.wikipedia.org/wiki/Denotational_semantics
[51] http://plato.stanford.edu/entries/dynamic-semantics/

node for exploratory data analysis. Whereas below we refer to scalability results demonstrated on texts and artworks as forms of semantic media, the approach is generic and applies to any features represented by vectors, including image descriptors. Its multipurpose nature means that it can be used to analyze concept and semantic change (the focus of Chapter 5) and to uncover hidden correlations in sparse data collections.

To connect the three key concepts in the subtitle, i.e. the acceleration of computational processes, evolution and the problem of digital object (DO) classification based on semantic content, we briefly refer to a key concept in classification by machine learning, that of **energy minimization**, as this leads to a new approach which considers semantics as "energy" in a metaphoric sense[52]. Based on the theory that energy is stored in fields in physics to induce change, we too can conceive semantics in a field form where the contextual nature of content is the "energy" driving minute changes whose sum total is underlying evolution. The proof for this line of thought is the scalability and generic applicability of Somoclu as a field analytical tool to the classification of evolving semantic media.

### *4.2.1. Testing Scalable Applicability to Text*

Having shown in D4.3 that Somoclu is significantly faster and more scalable than other approaches and available tools in this development area, before going into a detailed analysis of evolving semantics in Chapter 5, we carried out two initial tests on natural language text and artworks metadata, looking at a combination of scalability aspects and the semantic content of artefacts for a start. The first experiment was based on Stanford's Amazon book reviews data set as a collection of DOs [McAuley & Leskovec, 2013], which is publicly available as part of the University's SNAP project[53]. The data set spanned a period of 18 years and included approximately 12.8M book reviews up to March 2013. Every item in the data set included product and user information, ratings, as well as a plain text content description. We split the corpus in three periods, each containing close to 4.3M objects. The key characteristics are summarised in Table 4-1.

**Table 4-1**. Key statistics of the temporal split of the Amazon test corpus.

| | **Period 1** | **Period 2** | **Period 3** |
|---|---|---|---|
| **Date** | Until 30 Jan 2003 | Until 03 Aug 2008 | Until 04 Mar 2013 |
| **# of terms** | 45,162 | 49,400 | 50,672 |

The results were evaluated for their semantic consistency based on their statistical significance [Wittek et. al, 2015b]. It was decided that more research will be necessary to work out a comprehensive evaluation methodology to interpret the interplay of position and direction vectors that constitute a vector field. As the latter indicate emergent changes in the field, the dislocation of actual semantic content vs. potential displacements was to be addressed in terms of a dynamic theory of word semantics. This pointed in the direction of concept drifts and topic shifts as related research areas.

[52] Mathematical "energy" and machine learning (ML) are related, the latter often being based on minimizing a constrained multivariate function such as a loss function. Concepts in feature space "sit" at global energy minima, representing the cost of a classification decision as an energy minimizing process. This suggests that ML must identify concepts with such minima, and since energy in physics is carried by a field or a respective topological mapping, concepts naturally have something to do with energy as work capacity.
[53] http://snap.stanford.edu/

### 4.2.2. Testing Scalable Applicability to Artworks Metadata

We also tested Somoclu on the Tate art collection and archive holdings dataset publicly available for research. Details of this collection are listed in Section 5.3.1. The dataset contains approximately 69,000 records with the following metadata elements for each artwork and archive item:

- Artist(s)
- Title
- Date created
- Reference number / Accession number
- Medium description
- Web address (URL) of page in Art & artists section of the Tate website
- Subject index terms
- Image web address (URL)
- Credit line
- Movements

To get a first impression about the evolution of artistic technology between 1823-2013, we generated a 69,000 x 1,023 binary (presence-absence) matrix of artworks indexed by the medium used for self-expression. The results (see Fig. 4-1) indicated that in a next phase of research (within T4.5 and reported in Section 5.3.1), index terms from the hierarchical Tate subject index will be a suitable input to map trends of semantic content evolution.

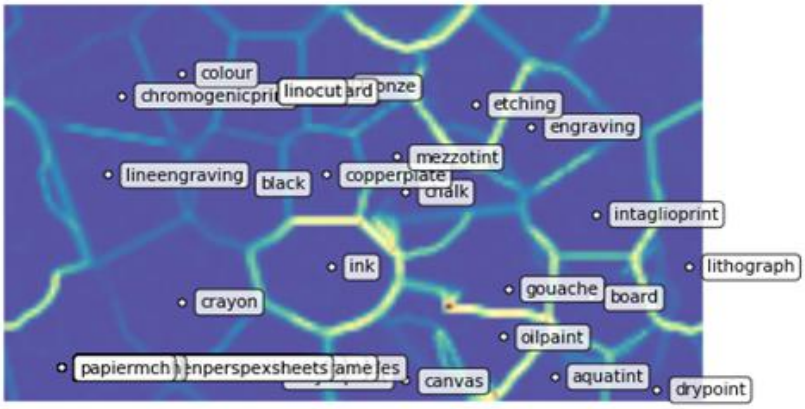


**Fig. 4-1**. A small (90 x 150 nodes) toroid map separating artistic media of self-expression over 190 years with boundaries indicating established against incoming trends. Partly due to the impact of the Turner bequest, old techniques are dominant.

## 4.3. Background Considerations for the Modelling of Semantic Evolution in a High Performance Computing Environment

To connect algorithm development for scalability with paving the way for modelling evolving semantics in Chapter 5, here we briefly revisit [Schlieder, 2010] to offer our background considerations. In his LTDP (Long-term Digital Preservation) problem typology, his second type of aging implied **language change**, whereas his third type referred to **cultural value modifications** turning old contexts obsolete, thereby making the interpretation of new words or new values in old contexts problematic.

For our approach to the above types of change in evolving Digital (Eco)systems, we combined the following observations:

- Societies, including human ones, exist because they categorize their percepts into classes based on similarities. A DP system, scalable or not, must preserve such categories to maintain the continuity of a particular society. A DP system of semantic content, evolving or not, has to provide full access to its holdings in the future, too.
- Computerized access to semantic content departs from information representation by mathematical objects for their comparison[54]. One particularly widespread convention uses vectors as the means of information (i.e. content) representation. Such vectors constitute vector spaces, practically geometries with two- to many-dimensional positioning of content in that space, suitable for information retrieval (IR) and machine learning (ML), based on the concept of similarity between objects and/or features. Whereas without ML, access to scalable systems of semantic content is hardly conceivable any more, the shortcoming of this representation convention is that vector spaces are stable, i.e. their formalism cannot model change.
- A vector-based model that includes both pointwise located content and its dislocations over time, is the concept of *fields* in classical mechanics (CM). There, a field is a continuous space in which value distributions of e.g. mass or electric charge characterize its regions. Evolving values are represented by location and direction vectors, change being described by differential equations in general and partial differential equations in multidimensional geometries in particular. One can generate a continuous model from a discrete vector space by interpolation.
- With context as the independent variable in vector space semantics based on the distributional hypothesis, the step toward adopting a vector field model to simulate evolving semantics is to consider the **theory of semantic fields** [Trier, 1934] as one particular theory of word meaning.
- Our selected methodology, ESOM, creates a vector field where content is located by context, evolves by displacement and redistribution, meaning complies with both Harris [Harris, 1968] and Trier, based on Aristotle who points beyond classical mechanics. The field is scalable; robust drifts exist in it, verifiable by different statistical methods, so that they can be detected, measured and interpreted.

To perceive the mechanics of the semantic drift, in such vector fields, content is fluctuating, i.e. it keeps on "flowing" from here to there, an infinite process that, time and again, results in new spatial distributions of different topical composition. All these go back to the evolving context of object features, such as the index terms characteristic for a set of DOs, adding up to a highly complex network of dependencies in progress. Thereby such "shapes" or morphologies of content characteristic for consecutive observation periods establish a sound point of departure.

With the above in mind, Somoclu was designed to act as a "telescope" for the observation of scalable, contextual, multivariate higher-order morphologies, including capabilities for the detection, measurement and interpretation of content drifts. This "telescope" realises the idea of information astronomy, inasmuch as it surveys the behaviour of any kind of content represented as an evolving vector field like an astronomical observatory of changes would[55]. Whereas in its current implementation, Somoclu uses word semantic content to characterize classes of DOs, a next extension will be able to group them by their sentence semantic content such as RDF statements, too. In other words, given scalable DP data in the future, such data can be indexed by LRM propositions and studied as an evolving field of dependencies if necessary (for more details see [Lagos & Vion-Dury, 2016]). Scalable collections of ontology-indexed DOs, on the other hand, open up new ways for *collection diagnostics* as a risk management approach over time.

[54] Such content can also be financial, emotional, functional, aesthetic, popular etc.

[55] The observatory concept can be extended e.g. to express any kind of content, including semantic content, by spectrograms [Wittek & Darányi, 2011]. Respective experiments are being conducted for D4.5.

## 4.4. Chapter Summary

In this chapter we departed from the notion of contextualised content semantics as a standard way of ascribing meaning to digital preservables, and presented a multivariate statistical approach to the scalable processing and accessing of such content. Also, we carried out initial tool tests to check the feasibility of our background considerations on a major text dataset and image metadata from the online catalogue of Tate Gallery. This will be the basis for the field approach to evolving semantics in Section 5.3.1 and for community change in Section 5.3.3 of the next chapter. At the same time Section 5.3.4, including guest research on topic shifts, will be focus-wise relevant but methodically different, using probabilities instead of vectors, but departing from the same statistical data.

# 5. Semantic Change and Evolving Semantics

This chapter focuses on the investigations and experiments we conducted with regards to studying semantic change and the overall phenomenon of evolving semantics. The chapter starts with a background survey of all relevant notions and terminology, followed by a brief discussion on the relevance of semantic change for DP, and then introduces the respective experiments we performed. Some interesting guest analytical work by partners outside PERICLES is also presented in this chapter.

## 5.1. Background

**Evolving semantics** (also often referred to as "**semantic change**") is an active and growing area of research that observes and measures the phenomenon of changes in the meaning of concepts within knowledge representation models, along with their potential replacement by other meanings over time. Therefore, it can have drastic consequences on the use of knowledge representation models in applications. Semantic change relates to various lines of research such as **ontology change**, **evolution**, **management** and **versioning** [Meroño-Peñuela et al., 2013]. It also entails ambiguous terms of slightly different meanings, such as **semantic shift** and **concept drift**.

**Table 5-1**. Terminology and aims of existing studies about semantic change. Each study targets one or more topics either directly or indirectly, marked with ✓ and (✓) respectively.

| Study, Year | Semantic Change | Semantic Drift | Concept Drift | Concept Shift | Concept Change | Concept Versioning | Topic Drift | Topic Shift | Semantic Decay |
|---|---|---|---|---|---|---|---|---|---|
| [Tury & Bieliková, 2006] | ✓ | | | | | | | | |
| [Wang et al., 2011] | | ✓ | ✓ | (✓) | | | | (✓) | |
| [Fanizzi, et al., 2008] | | | ✓ | | (✓) | | | | |
| [Uschold, 2000] | | | | | ✓ | | | | |
| [Meroño-Peñuela et al., 2013] | | | ✓ | | (✓) | | | | |
| [Wang et al., 2009] | | | (✓) | ✓ | (✓) | | (✓) | | |
| [Wittek et al., 2015b] | | ✓ | ✓ | | (✓) | | | | |
| [Yildiz, 2006] | | | | | (✓) | (✓) | | | |
| [Klein & Fensel, 2001] | ✓ | | | | | (✓) | | | |
| [Gulla et al., 2010] | (✓) | ✓ | | | | | | | |
| [Stojanovic et al., 2002] | | | | ✓ | | | | | |
| [Klarman et al., 2008] | | | | | | ✓ | | | |
| [Pareti et al., 2015] | | | | | | | | | ✓ |

This section presents a background study overview to disambiguate the different terms within the semantic change research area. Table 5-1 presents an overview of all studies related to each one of the different terms and fields. As many studies use the terms interchangeably as synonyms, and their differences are subtle, in this overview we consider each study to be targeted at one or more fields of semantic change, either directly or indirectly; studies directly target a field when they give definitions and present methods to measure and investigate it, and indirectly when they consider it a secondary target, or simply a synonymous term for their main, direct target.

These emerging relationships between the terms are given in more elaborate detail in Fig. 5-1. The graph shows a large node for each term and field within semantic change, linked with studies that directly refer to it. The studies are further linked with terms they target indirectly (in grey). Overall, *concept drift*, *concept change* and *concept shift* are closely related, but still connected to *semantic drift*, while *semantic decay* remains a self-contained field of study.

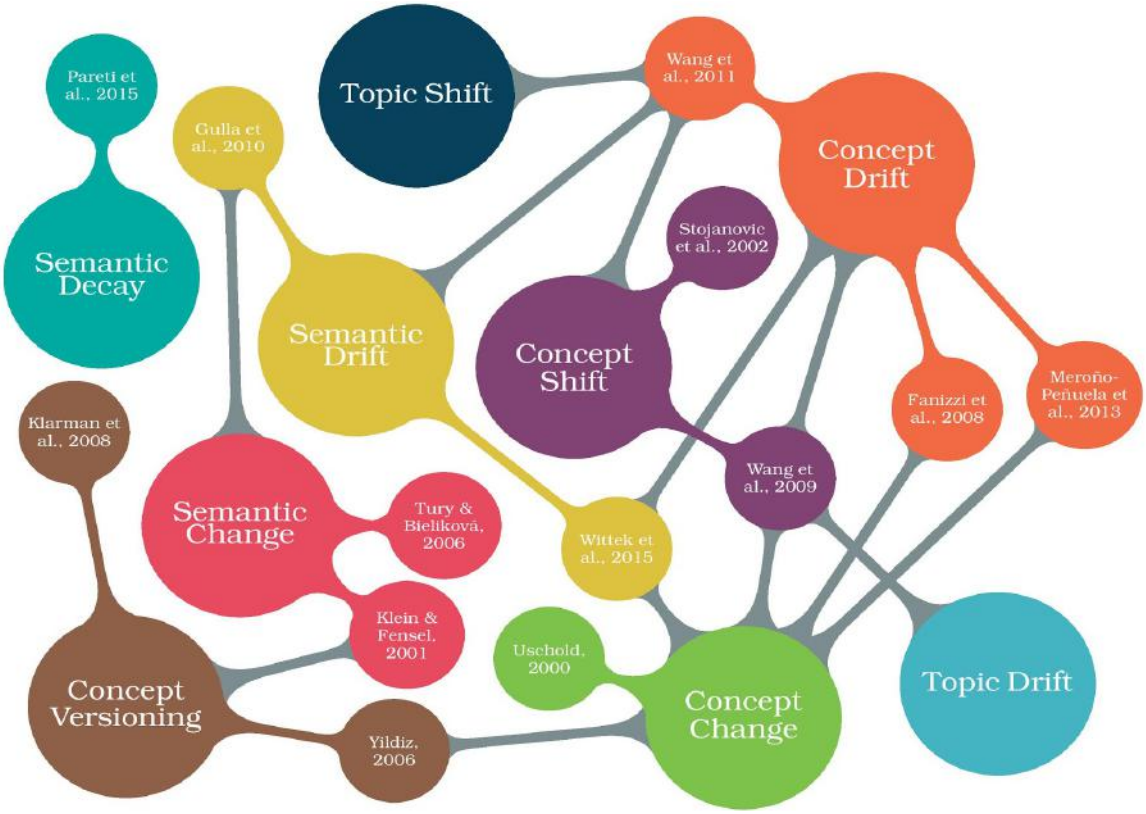


**Fig. 5-1**. Relationships between terms of semantic change through literature.

In Fig. 5-1, not only the relationships between fields are apparent, but also their penetration and adoption. For instance, the most popular terms and fields are *concept change*, *concept drift* and *concept shift*, with the least popular being *semantic decay*, *topic shift* and *topic drift*. However, it should be noted that some of the terms (e.g. semantic decay) seem to have been introduced only very recently when compared to others.

Further details for each term, including definitions and methods to track and measure semantic change from existing research are given below.

**Semantic Change**

Semantic change refers to the extensive revisions to a single ontology or differences between two ontologies and, therefore, can also be associated to versioning. According to [Tury & Bieliková, 2006], semantic change occurs when the internal structure of a concept in two ontologies is different. On

the contrary, an isomorphic change refers to the structure being unchanged while the names might have been altered. [Klein & Fensel, 2001] consider semantic change when an ontology revision presents so many alterations that it can be reformed as a new conceptualization, with its own identity. The ontology authors are responsible to decide whether semantic change will occur or the ontology will continue to represent the same conceptualization.

**Semantic Drift**

Semantic drift refers to how the features of ontology concepts gradually change as their knowledge domain evolves, or, alternatively, how they can be reinterpreted by different user communities in a different context, introducing a risk for them to lose their rhetorical, descriptive and applicative power [Wittek et al., 2015b]. It can also be defined as the gradual change of a concept's semantic value, as it is perceived by a community. It can be characterized as intrinsic or extrinsic, depending on whether a concept's semantic value is changed with respect to other concepts in the ontology or to the phenomena it describes in the real world.

Drift can also be classified as non-collective, inconsistent or consistent collective [Gulla et al., 2010]. If a concept is exposed to extrinsic, but no intrinsic drift, it means that the whole ontology is undergoing a collective consistent drift that may not necessitate any changes to the ontology. On the other hand, no extrinsic drift and substantial intrinsic drift means that a concept's relationships to other concepts in the ontology may no longer be correct, even though the concept itself has not changed its meaning. In cases of both extrinsic and intrinsic drift we are dealing with inconsistent collective drift of concepts in an ontology that is no longer valid.

**Concept Drift**

Generally, concept drift is defined as a change in the meaning of a concept over time, but possibly also over location, culture, etc. It often refers to a problem in the field of data mining or machine learning, when learned models lose their predictive power over time [Wang et al., 2011]. On the other hand, [Wittek et al., 2015b] entirely separate the two terms and fields. They define concept drift as the abrupt parameter value changes that occur in data mining, while semantic drift is the language-related version of the same phenomenon. However, [Wang et al., 2011] has bridged the gap by applying notions from concept drift in data mining, such as intension, extension and labelling, as means for measuring semantic drift as well.

Concept drift or topic drift [Wang et al., 2009] can also be defined as the change of known concepts based on evidence provided by new annotations that emerge over time. In that sense, concept drift can be detected by investigating alterations of concept clusters over time, formed according to various similarity criteria [Fanizzi et al., 2008]. In [Meroño-Peñuela et al., 2013], three types of drift are discriminated: concept label, intensional or extensional drift (i.e. a change of meaning that affects the extension of a concept).

**Concept Shift**

Concept shift refers to the subtle changes in meaning of related concepts over time. It can be studied by using chains of extensional, i.e. instance-based, mapping that represent those subtle changes [Wang et al., 2009]. Concept shift often occurs in the course of evolution so that the actual meaning of concepts better represent the structure of the real world. While some shifts of concept meaning are performed explicitly, they can also be implicit, through changes in other parts of the ontology, e.g. in properties [Stojanovic et al., 2002]. The term "topic shift" can also describe the same phenomenon [Wang et al., 2011].

**Concept Change**

Concept change refers to the broad variety of adaptations and alterations that can occur for a concept in an ontology. Such changes can be either conceptual (e.g. changing concept relations), specification or representation [Yildiz, 2006]. In [Uschold, 2000] advises on how to track concept change by investigating obsolete concepts that have changed name, but maintained their identifiers and history of changes that can later be examined.

**Concept Versioning**

Concept, or more generally, ontology versioning refers to building, managing and providing access to different versions of an ontology [Yildiz, 2006]. Another definition is that versioning methodology provides users of the ontology variants with a mechanism to disambiguate the interpretation of concepts [Klein & Fensel, 2001].

It is also linked (but not identical) to ontology evolution by the ontology and database engineering communities, as both research fields aim to represent change and handle different variants of ontologies [Yildiz, 2006]. However, one of the differences between them is that ontology evolution concerns changing an existing ontology while maintaining consistency, whereas ontology versioning follows a copy-first strategy where changes are effected in a new, duplicate version of an ontology [Klarman et al., 2008].

**Semantic Decay**

Semantic decay refers to the declination of semantic richness of concepts. The amount of facts that can be inferred from a concept, within the context of Linked Data and a particular dataset, has been proposed as a measure of richness and thus semantic decay. Using this metric it has also been proved that the more a concept is reused, the less semantically rich it becomes [Pareti et al., 2015].

## 5.2. Semantic Change and Digital Preservation

There is a need to tackle the challenge of semantic change within the DP setting. One can gradually lose the cultural heritage of a civilization in different ways. Following Schlieder's view on risks affecting LTDP [Schlieder, 2010], a particular kind of erosion is due to inevitable **technological changes**, with hardware and software obsolescence turning what used to be computer-readable into corrupted bits, e.g. due to "bit rot". Another source of danger is **language change**: with progress, new concepts have to be named and old ones' meanings shift in new directions either in the population as a whole, or in pockets of different uses. Thirdly, due to emerging **social pressures**, cultural values also undergo unpredictable modifications, so that e.g. what used to be forbidden yesterday may become compulsory tomorrow. Due to these three streams of intertwined dynamics, DP is facing a difficult situation.

Whereas PERICLES' mission, among others, is to call attention to all three kinds of potential risks, here **we focus on the nature of semantic change, how it can be detected, measured, interpreted and remedied in digitized material**. As the task is complex and its scalable handling is only about to start, no final solutions, only intelligent choices can be indicated.

The problem of semantics in culture goes back at least two millennia. Asking for the meaning of sentences and words had started in Vedic India, continued in classical Greece, and through scholasticism and different philosophical undercurrents, culminated in a great number of theories of word and sentence semantics by the 20$^{th}$ century. This alone hints at the fact that no single, unified theory of meaning exists up till now. Information theory encouraged people to believe that they understand the problem because they can mechanize the communication process, although Claude Shannon made it clear that information theory left out semantics and dealt with communication on a formal ground only [Shannon, 1948]. Worse, the very term "information retrieval" from the sixties

onward reinforced misunderstandings and masked the fact that people are interested in meaningful answers by the computer, whereas the nature of semantics is still cryptic.

An early warning to the information retrieval community was the problem of inter-indexer consistency, subject to renewed interest in the WWW environment [Chi, 2015], as follows. Given a test set of documents to be manually assigned keywords to signal their content, human indexers could not agree between themselves what the subject of a certain document might be and indexed it by only partly overlapping keywords. Automatic indexing [Luhn, 1957] replaced these insecurities by statistical assumptions but, as the history of evaluation in information retrieval and text categorization over the past fifty years has shown, a solution to scalability did not answer the original question. In all, inconsistent indexing limits future access to DOs of any kind, and fluctuations in word meaning due to changes in word use or by a need to redefine concepts pose such a constant danger.

## 5.3. Adopted Investigations

The following subsections address our proposed approaches addressing the problem of detecting, measuring and interpreting semantic change in three different directions:

- **Field approach to evolving semantics**, dealing with textual content and terminology change, discussed in Section 5.3.1;
- **Semantic change under an ontology evolution perspective**, dealing with changes occurring in ontology models, discussed in Section 5.3.2;
- **Studying community change** in social media, discussed in Section 5.3.3.
- Finally, Section 5.3.4 presents analytical work on **topic shifts** conducted by partners outside the PERICLES consortium.

### *5.3.1. Field Approach to Evolving Semantics*

In order to combine semantics from computational linguistics with evolution, we selected the **theory of semantic fields** [Trier, 1934] and blended it with **multivariate statistics** plus the **concept of fields in classical mechanics** to enable machine learning. Semantic fields reproduced by statistics are context-based, therefore comply with the basic stance of PERICLES. Fields in physics are evolving, thus prove to be a suitable model of evolving semantics. The anticipated outcome of the planned experiment was to verify the following working hypothesis:

- The 2-dimensional surfaces that represent the indexing vocabulary of a collection at a point in time change shape, much like a landscape would under pressure from tectonic forces;
- Changes in the field manifest themselves as content drifts, tracked down and analyzed by term cluster consistency checks.

#### *Tool Design for Working Hypothesis Testing*

In the DoW, contextualised semantics refers to the variations in meaning and interpretation that arise according to the context, in which content is taken into account to develop a model that accommodates this context-dependent nature of content semantics. To address issues related to the modelling of semantic evolution on accelerator-enabled hardware, we departed from the following considerations:

- Evolving semantics manifests itself in the changing topical composition of the collection, expressed by the features of the objects;
- Drifts in feature values have an impact on access, and thereby the returns of DP as an investment - with limited access, DP would fail its ultimate goal;

- A monitoring tool with drift detection, measurement and semiautomatic interpretation capabilities can help digital curators by issuing, e.g., a system alert given a threshold of semantic changes in terms of drifts.

We designed a proof-of-concept tool based on the metaphor of **information cosmology**, modelling evolving semantics on the field nature of an expanding information universe with thematic galaxies in it [Olsen et al., 1993; Wise 1999]. This observatory-like tool anticipated micro- to macroscopic analytical abilities, i.e. zooming in and out, and took snapshots of content distributions at regular intervals to record changes. We were interested in the workflow and scalability aspects of both its in- and output.

The tool can scan scalable sets of objects and/or features to inspect trends on the deepest level given by e.g. an index term hierarchy. Here we present findings from a test on language. It is obvious that in an extended version, not just words but phrases or sentences could be used for indexing as well, or any features describing any objects for that matter, therefore the tool is generic.

### *Material and Method*

To test the observatory concept, we opted for the Tate catalog public metadata as a step toward scalability. The total number of records was 69,202, out of which the 53,698 were timestamped. The records are JSON formatted, which makes them easier to read and parse, since the aforementioned format is included in a wide range of libraries and is reasonably succinct.

Based on statistical sampling, two 50-year periods (1796-1845, 1960-2009) were selected (Fig. 5-2, Fig. 5-3) each of which were divided into 10 x 5-year epochs. In these two periods, 46,381 records constituted the sample in analysis, with 33,625 artworks between 1796-1845, and 12,756 between 1960-2009.

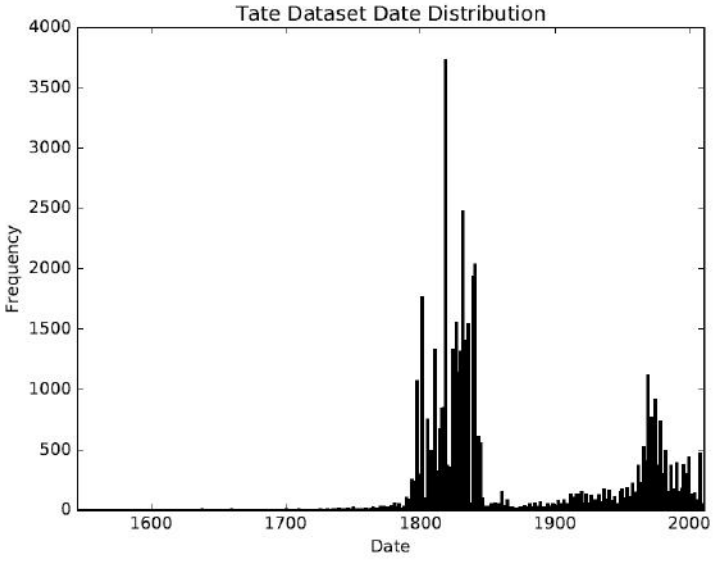


**Fig. 5-2**. Acquisitions for the Tate collection in chronological order.

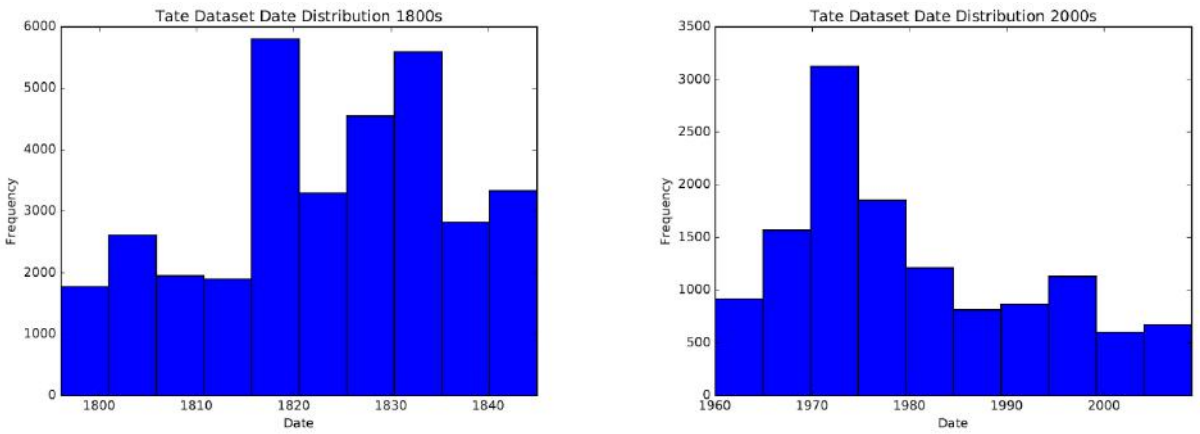


**Fig. 5-3**. Acquisition rates in the Tate collection between 1796-1845 and 1960-2009.

The Tate catalog metadata are indexed hierarchically on three conceptual levels but not by a thesaurus or a formally defined ontology:

- In the first period (1796-1845):
  - Level 1 has 22 unique terms (21 of which are persistent, i.e. present over all timesteps/epochs),
  - Level 2 has 203 unique terms (142 of them persistent), and
  - Level 3 has 6,624 unique terms (225 of them persistent).
- In the second period (1960-2009):
  - Level 1 has 24 unique terms (22 persistent ones),
  - Level 2 has 211 unique terms (177 persistent ones), whereas
  - Level 3 has 7,536 unique terms (288 of which are persistent).

It is immediately apparent from these numbers that on the most detailed description level the terminology is highly volatile, with only 3.4% and 3.8% respectively of the terms being persistently applied to the incoming objects. Fig. 5-4 shows a sample subject index entry from the catalog.

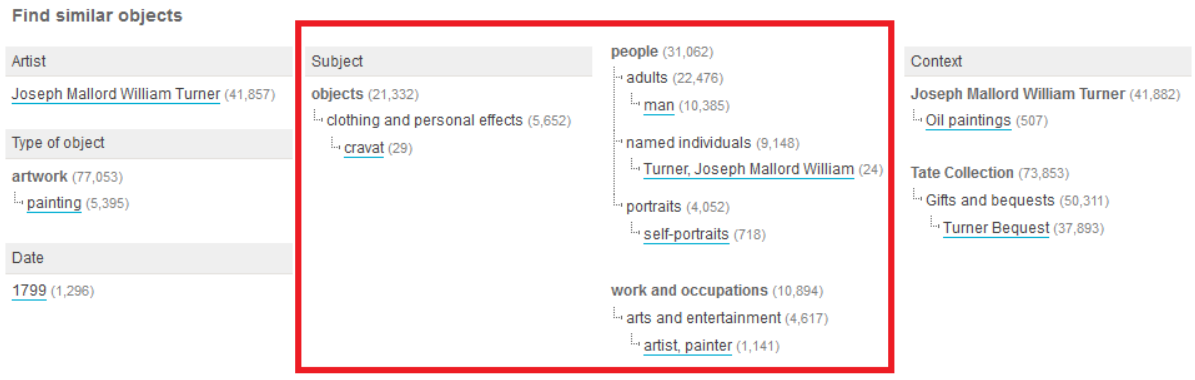


**Fig. 5-4**. Sample entry in the Tate subject index.

To test the field idea, we wanted to evaluate term clusters based on their **semantic consistency** [Wittek et al., 2015b]. To this end, we opted to use the UCREL semantic tagset[56] used for corpus analysis by the Wmatrix tagger [Rayson, 2008] as an intermediary step. UCREL groups terms into lexical fields, so semantic tagging meant that we indexed the artefacts by these. The mapping relied

[56] http://ucrel.lancs.ac.uk/usas/

on the respective algorithm of Wmatrix and was not cross-validated. An example of how Tate subject index terms correspond to UCREL tags and category labels is shown in Table 5-2.

**Table 5-2**. Correspondence between Tate subject index terms and UCREL categories.

| Tate Term | UCREL Code | UCREL category |
|---|---|---|
| abbey | H1 | Architecture, houses and buildings |
| abstract | A1.6 | Concrete/Abstract |
| abstraction | A1.6 | Concrete/Abstract |
| actions | A1.1.1 | General actions / making |
| activities | A1.1.1 | General actions / making |
| actor | K4 | Drama, the theatre and show business |
| adults | T3+ | Time: Old; grown-up |
| advertising | I2.2 | Business: Selling |
| aggression | E3- | Violent/Angry |
| agricultural | F4 | Farming & Horticulture |
| agriculture | F4 | Farming & Horticulture |
| air | O1.3 | Substances and materials: Gas |
| aircraft | M5 | Flying and aircraft |
| alps | Z2 | Geographical names |
| ambiguity | Q3 | Language, speech and grammar |
| angel | S9 | Religion and the supernatural |
| animal | L2 | Living creatures: animals, birds, etc. |
| animals | L2 | Living creatures: animals, birds, etc. |
| anxiety | E6- | Worry |
| appliances | O2 | Objects generally |
| aquatic | M4 | Sailing, swimming, etc. |
| arch | H2 | Parts of buildings |
| architectural | H1 | Architecture, houses and buildings |
| architecture | H1 | Architecture, houses and buildings |
| aristocrat | S7.1+ | In power |
| ... | ... | ... |

### *Experiment Workflow*

A workflow for the following experiment design was established to study term-term (feature-feature) vs. term-document (feature-object) correlation matrices, with the emergence of Trier's word semantic fields by vector fields as the target:

1. Text processing: only terms indexing artefacts were included, but not their captions or abstracts. On this basis, for *term* x *term* matrices, index terms were replaced by UCREL codes using Wmatrix for semantic tagging. This resulted in three kinds of input matrices for statistical analysis: artefacts described by Tate index terms; by corresponding UCREL codes; and by labels for the respective UCREL codes. For *term* x *document* matrices, UCREL encoding was not tested;
2. As the Tate subject index is hierarchical with three levels of content description, for the above three kinds of input, matrices on 7 levels of granularity were generated (20x12 = level 1 [lvl1], 40x24 = level 2 [lvl2], 50x30 = level 3 [lvl3], 60x40 = all levels together [lvlA], plus resolutions of 100x60, 150x90, 200x120 grid nodes for zooming in);

3. The input matrices were processed by ESOM combined with affinity propagation clustering by Somoclu. The results were tested for robustness by hierarchical cluster analysis (HCA), using Euclidean distance as similarity measure and farthest neighbour (complete) linkage to maximize distance between clusters, keeping them thereby both distinct and coherent. The ESOM-based cluster maps expressed the evolving semantics of the Tate collection as a series of 2-dimensional landscapes over ten epochs per two periods. These went back to the two acquisition peaks between 1796-1845 and 1960-2009, each of these 50 years periods having been separated into ten 5-years epochs, respectively;
4. Term drift detection, measurement and interpretation were based on these maps. To enable drift measurement, we generated a parallel set of input matrices with the term of greatest PageRank centrality over all periods as its "Greenwich" point. This relative location was used as the anchor for the computation of respective term-term distance matrices over every epoch of each period. Finally, term dislocations over epochs were logged, recording both the splits of term clusters mapped on a single grid node in a previous epoch, or the merger of two formally independent nodes labelled with terms into a single one.

**Technical Details of Tool Design**

The task of drift detection, measurement and interpretation is carried out in three basic steps as follows:

- **Step 1**: Somoclu maps the high-dimensional topology of multivariate data to a low-dimensional (2D) embedding by ESOM. The algorithm is initialized by Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), or random indexing (RIX), and creates a vector field over a rectangular grid of nodes of an artificial neural network (ANN), adding continuity by interpolation among grid nodes. Due to this interpolation, content is mapped onto those nodes of the ANN that represent best matching units (BMUs), and are located in basins with ridges around them. Content splitting tendencies are indicated by the ridge wall width and height around such basins. Consequently, the ESOM method yields an overlay of two aligned contour maps in change, i.e. actual content structure vs. actual tension structure.
- **Step 2**: Clustering over this low-dimensional topology marks up the cluster boundaries to which cells with BMUs belong. The clusters are located within ridges/watersheds [Ultsch, 2005; Tosi et al., 2014; Lötsch & Ultzsch 2014]. In Somoclu, nine clustering methods are available. Self-organizing maps (SOM), including ESOM, reproduce the local but not the global topology of data (i.e. the clusters are locally meaningful and should be consistent on a neighbourhood level only). For mathematical reasons, Somoclu combines this feature with a toroid representation, i.e. content is mapped onto the surface of a doughnut (torus, a ring made of a tube) whose left and right vs. lower and upper boundaries connect. Fig. 5-5 shows a toroid map with three distinct basins for related content but the *K*-means clustering algorithm set to $K = 5$ clusters. To this end, the codebook was initialized before training with the PCA subspace of the first two eigenvalues, a procedure that results in the best of two worlds: a globally optimal embedding adjusted to reflect the local topology, so that thereby previously misclassified points vanish.
- **Step 3**: Evolving cluster interpretation by semantic consistency check. This can be measured relative to an anchor (non-shifting) term used as the origin of the 2-d coordinate system, or by distance changes from a cluster centroid, etc. In parallel, to support semiautomatic evaluation, variable cluster content can be expressed for comparison by histograms, pie diagrams, etc., of semantically tagged cluster content.

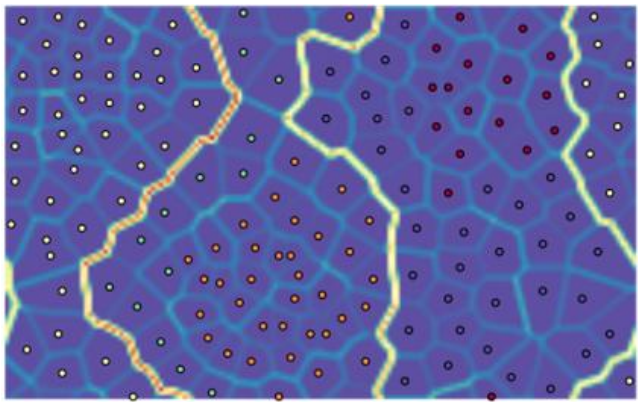

**Fig. 5-5**. A toroid ESOM map with three distinct basins for related content but the K-means clustering algorithm set to K = 5 clusters.

### *RESULTS*

The above yielded 1,600 2D maps (landscapes) as the core of our term-term correlation analysis combining *4 Tate subject index levels* x *3 codes* x *10 epochs* x *2 periods* x *7 granularity levels* x *2 measurement variants* (with and without anchor term). Another 240 maps were induced to study term correlations from the term-document matrices. Drift detection, measurement and evaluation were based on their analysis, leading to 560 drift logs on all indexing levels for both term-term and term-document analysis. In parallel, covering every timestep of collection development, we extracted altogether 168 normalized histograms to describe the evolving topical composition of the collection, and 320 pie charts describing the thematic composition of the clusters, both in terms of UCREL tags. Further, for Somoclu cluster robustness check, 80 HCA dendrograms were computed for term-term matrices and another 60 for term-document matrices (lvlA was disregarded here). The total number of figures generated was 3,924.

A detailed report of our complete analysis would go beyond the opportunities of this deliverable, but a related publication with the main findings is in progress. However, some key indications were as follows:

- The capacities of the method make the observation of coordinated evolution between content vs. tension structure possible. Content mapping means that term membership for every cluster in every timestep is recorded; term positions and dislocations over timesteps with regard to an anchor position are computed, thereby recording the evolving distance structure of indexing terminology. This amounts to drift detection and its exact measurement. Adding a drift log results in extracted lists of index terms on all indexing hierarchy levels plus their percentage contrasted with the totals. The evolution of the tension structure is documented, too. This signals the tendencies of conceptual splits and merges, affecting access to DOs. Here, *tendency means a projected possible, but not necessarily continuous, trend* - should the composition of the collection continue to evolve over the next epoch like it used to develop over the past one, the indicated splits and merges would be more probable to form new content agglomerations than random ones.
- To validate our clustering methods by the semantic consistency of their results, we estimated and compared the average path-based semantic similarity between clusters of terms formed by and/or over the neural network vs. randomly distributed ones. Path similarity denotes how similar two word senses in a cluster are, based on the shortest path that connects the same senses in the “is-a” taxonomy [Rada et al., 1989] e.g. in WordNet [Fellbaum, 1998]. In particular, we studied both the proximity of terms on the toroid plane (terms coinciding on the same neuron and forming a cluster, called BMU clustering), and terms whose weight vectors were clustered together by the affinity propagation algorithm. For semantic consistency evaluation,

clusters were formed by randomly distributing the terms into series of similarly sized containers as the ones constructed by the methods mentioned above. As an example, for the lvl2 persistent indexing vocabularies of the 2000s series, a comparison of the average similarity distribution for each of the methods is presented in Fig. 5-6. As the red and blue lines indicate, semantic similarity is clearly higher in the clusters formed by any of the ESOM-based methods than in the randomly distributed ones. It should also be noted that the clusters formed on the toroid are significantly fewer in number and more concise than the ones in the weight vector space which explains the difference in average similarity between them.

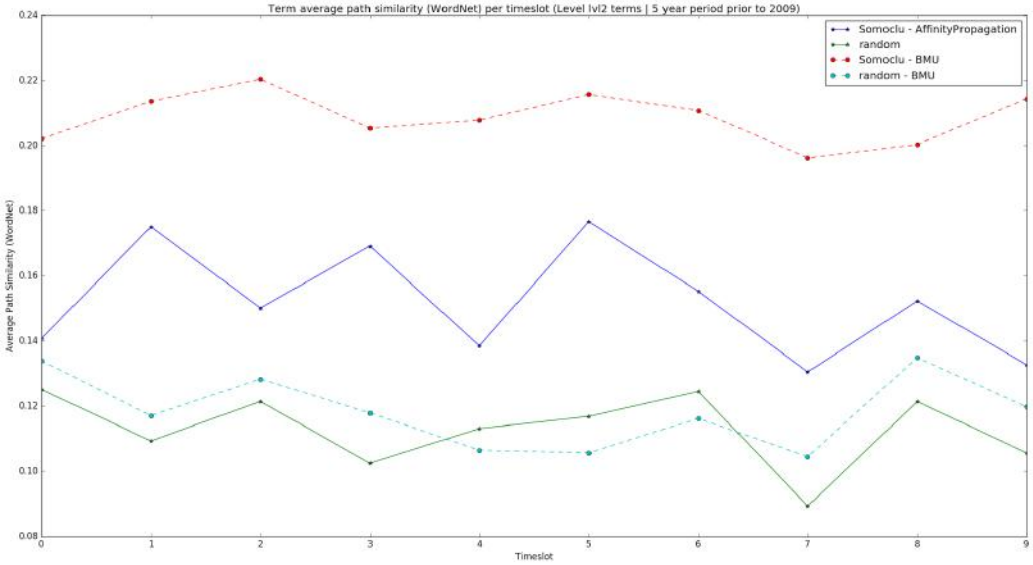


**Fig. 5-6**. Semantic consistency of level 2 term clusters in 2005-2009 by the average similarity distribution for four methods.

- All the terms and their respective semantic tags are in constant flux due to external social pressure, e.g. by novelties in the collection due to the composition of donations, fashion, etc. Without data about these pressures quasi embedding and shaping the Tate collection, the correlations between social factors and semantic composition of the collection cannot be explicitly computed and named. Any future modelling effort will need both components, not just content alone. However, some trends can be visually recognized over both series of maps, which is encouraging. These trends go back to the relatively constant semantic structure of the maps where temporary content dislocations do not seriously disturb the *relationships* between terms, i.e. neighbours tend to remain neighbours. This speaks for the relative stability of lexical fields as locally represented by Somoclu. E.g. in this particular collection, in spite of the high drift rate (at 40x24 resolution, i.e. for level 2 index terms, it was 19-22% for the 1796-1845 period, see Fig. 5-11 as well), terms clustered in the same attractor basin for the 1796-1800 epoch (such as “towns, cities, villages” or “uk, contries”) reappear in the 1801-1805 epoch too, regardless of context-induced changes in the overall landscape (see Fig. 5-7).

  To describe the composition of the social tensions shaping the collection, we compare the lvl2 persistent indexing vocabularies of the 1800s vs 2000s series. This is where we can witness the workings of language change, part producing new concepts, part letting certain ones decay. E.g. it is fascinating to see how new concepts emerge, i.e. how focus is shifting from a concept to its variant (e.g. *nation* to *nationality*), a renaissance of interest in the transcendent beyond traditional concepts of religion and the supernatural (*occultism*, *magic*, *tales*), fascination for the new instead of the old, or a loss of interest in *royalty* and *rank*. *Toys a*nd concepts like *tradition*, the *world*, *culture*, *education*, *films, games*, *electricity* and *appliances* make a debut in art. The

sum total of such appearances and disappearances goes back to the aforementioned social tensions, manifest in the impact they exert on the composition of the indexing vocabulary. Their comprehensive name is *progress*, but this label must remain tentative until better founded scientific results become available.

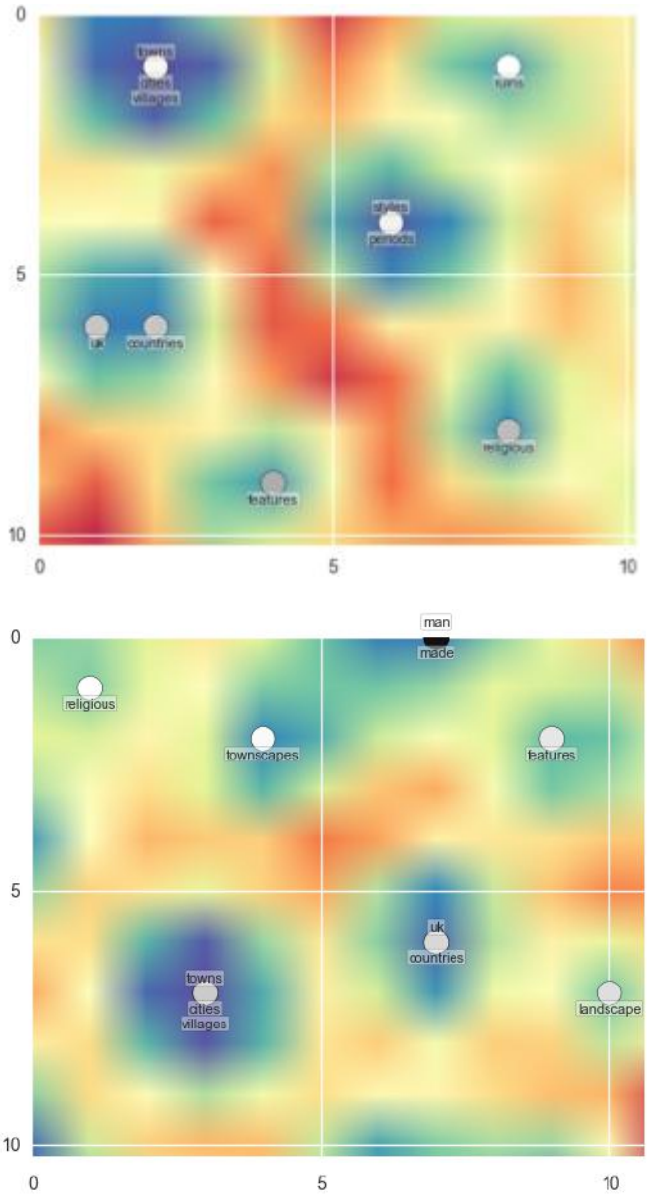


**Fig. 5-7**. **Upper picture**: level 2 index terms in the 1796-1800 epoch in the x = 0-10; y = 0-10 segment of the ESOM map. **Lower picture**: the same map area between 1801-1805. In spite of the 19-22 % average semantic drift rate between 1796-1845, the lexical fields with "towns, cities, villages" or "uk, contries" survived context-induced changes.

- With word meaning typically represented as locations in vector space, we know the interpretation of the vectors pointing at them. For every term vector, its interpretation is the respective word meaning in vector space used as its label.

  On the other hand, if for the detection and measurement of semantic drift we must apply a vector field model, one that comes with location plus direction vectors, we need to interpret the meaning of the latter kind as well. Ultimately the question is, *what does the direction of content displacement refer to*? This is work in progress and subject to future research by a much broader

community than the subject areas represented in PERICLES. The closest avenues to explore in linguistics and language philosophy are *update semantics* and *dynamic semantics*, and what will have to be found is to interpret tension structure in a vocabulary in terms of language change.

As [Veltman, 1996] defines update semantics, "The slogan '*You know the meaning of a sentence if you know the conditions under which it is true*' is replaced by this one: '*You know the meaning of a sentence if you know the change it brings about in the information state of anyone who accepts the news conveyed by it*'. Thus, meaning becomes a dynamic notion: the meaning of a sentence is an operation on information states." It's enough to replace "sentence" by "word" in the above definition to arrive at words as labels of single concepts inducing change, in line with Bloomfield's famous definition of word meaning, namely that the meaning of a word is its consequences.

With this caveat, we refer back to the fact that a vector field represented by ESOM is the composite of a content structure and a tension structure. We can measure changes in the compositionality and semantic consistence of the content structure over time (e.g. Fig. 5-8 in next subsection), plus in principle, estimate the likelihood of changes by splitting tendencies in the tension structure; however currently the latter falls short of prediction.

### *Discussion*

We have resolved **drift detection** and **drift measurement**, and partly resolved **drift interpretation**, with the automatic evaluation of cluster consistency accomplished.

For the detection task, our detailed and thoroughly documented findings indicate that in an evolving collection, **drift is the norm, not the exception**. Apart from surveying the evolving content structure, ESOM by Somoclu also scans the parallel evolution of classification tension structure, a precondition to future modelling and anomaly prediction.

As for drift measurement, we worked out a method to pin down the conceptual origin of the 2-d maps over the analytic periods, so term dislocation can be quantified with respect to it. For lvl1, in the first period, this "conceptual North Pole" was the term *nature*, for the second one, *concepts*. For lvl2, in the first period, the anchor term was *UK*, in the second period, *qualities*. For lvl3, the respective anchor terms were *river* and *man*, whereas for lvlA, they were the *UK* and *concepts*, respectively. As these anchor terms were relatively the most stable ones, i.e. the least drifting in an actively changing content landscape, all terms drifts were measured with respect to them.

Regarding drift interpretation, we used semantic tagging to express the composition of evolving term clusters, and plotted them as pie charts with UCREL tags against the normalized histograms for every epoch. Such a chart reveals important insights about the topic which a group of related artefacts happen to manifest and is beyond single concepts listed in dictionaries. The difficulty with the correct automated interpretation of such output is that for reasons of contextuality, to be able to refer back to the social context regulating our test data, one would need to have such embedding data as well. Such a specific test collection with two layers where content is dependent on context for interpretation studies is currently missing.

Due to our drift monitoring method, **a digital curation tool on an unprecedented scale and in unrivalled detail is now available**.

Work with the content structure of the Tate catalog has confirmed the working hypothesis, because splits between level 1 concepts took place occasionally, whereas both splits and merges occurred on levels 2-3 on a regular basis. From an information retrieval perspective, splits decrease recall, while merges decrease precision, limiting access. This is illustrated by Fig. 5-8 with splitting terminology between 1964-1974 at resolution level 20x12 (lvl1).

In 1964, the locations of *concept* and *emotion* as index terms applied to DOs overlap and the intensively coloured cell wall indicates a splitting tendency from the location of *idea*. In 1969, the

split of *idea* is complete and *concept* and *emotion* also start splitting, their separation being complete by 1974.

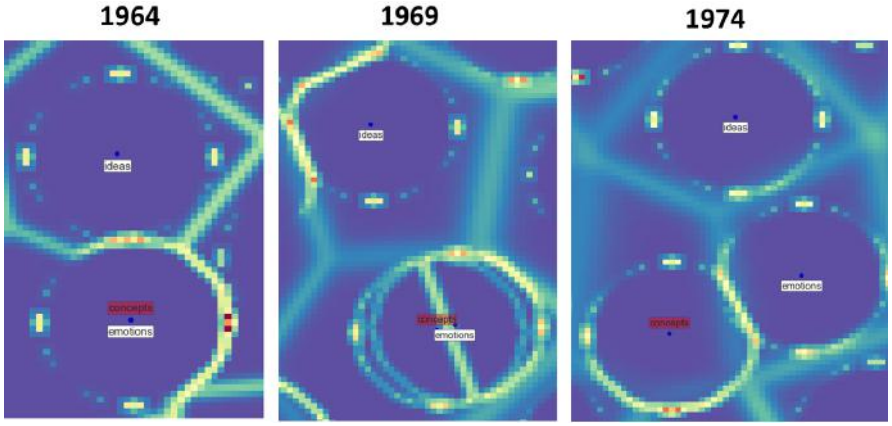


**Fig. 5-8**. Somoclu output of splitting level 1 terminology.

The drifts, separated into splits and merges, are listed for every epoch over both periods. For instance the file `changes1800slvl2_1.txt` for resolution 200x120 states the following:

```
Terms art at 174,88 were split from 184,101
Terms scientific at 138,68 were split from 11,113
Terms monuments,places,workspaces at 2,65 were merged from 1,73|20,30|13,97
Terms measuring at 17,66 were split from 11,113
Terms works at 183,86 were split from 184,101
```

This means that due to new entries in the catalog between 1796-1800, by 1800 on subject index level 2 the term *art* was split from *works*, just like *scientific* from *measuring*, whereas *monuments*, *places* and *workspaces* were merged, i.e. were mapped onto the same coordinates. Therefore, based on the same subject index terms, anyone using this tool in 1800 would have been unable to retrieve the same objects as in 1796 (In the above example, e.g. `184,101` indicates grid coordinates x=184, y=101). For more details, see Fig. 5-9 and Fig. 5-10. (As term labels are not readable in the printed version, we recommend the reader to study them by zooming in on figures in the original D4.4 file).

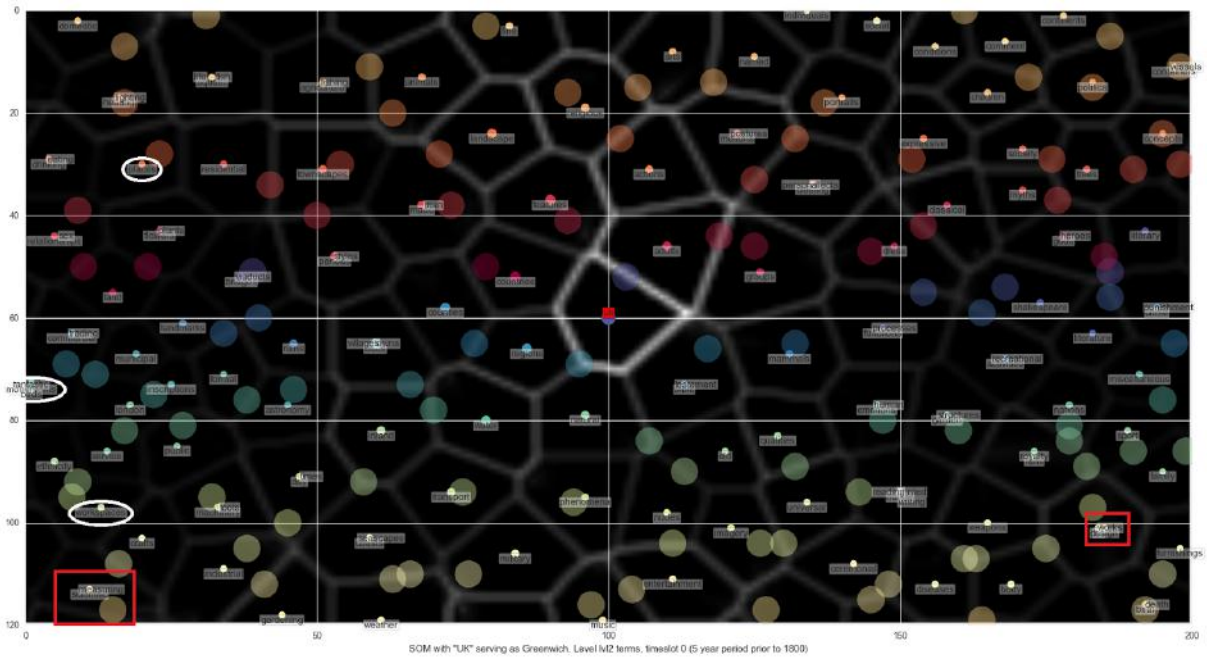


**Fig. 5-9**. In the first level 2 snapshot at 1796, the white ovals contain monuments, places and workspaces split as separate concepts. The red box to the left displays scientific and measuring merged, just like the orange box with art merged with works into a single concept.

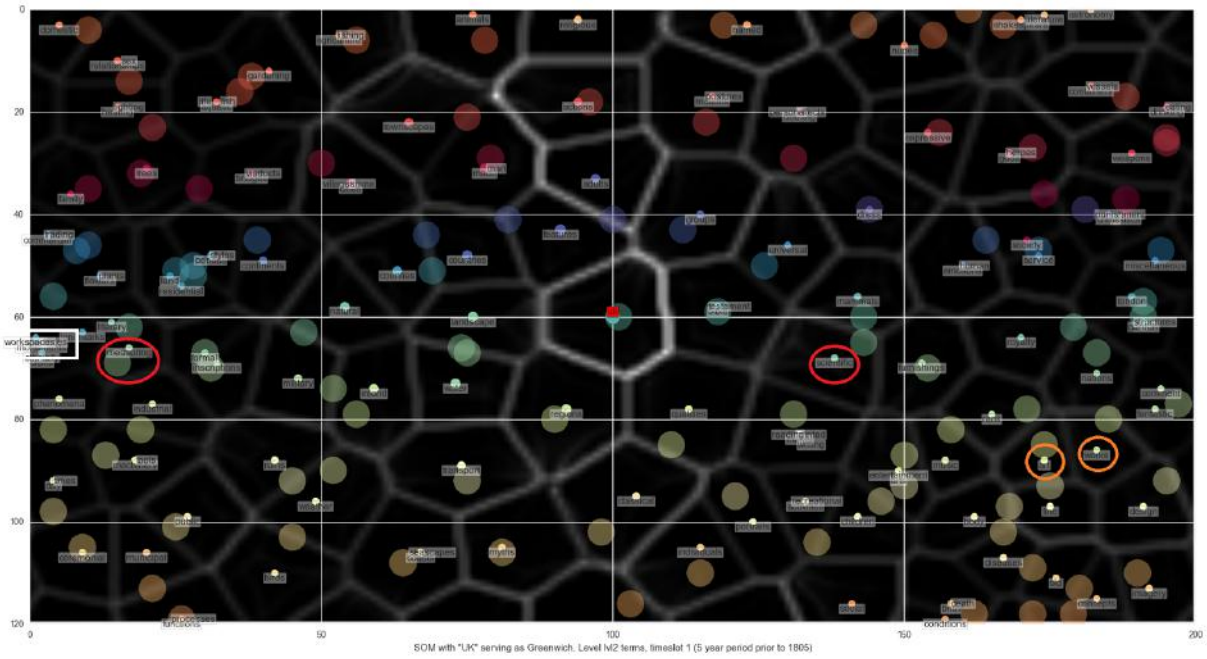


**Fig. 5-10**. In the second level 2 snapshot at 1800, the white box contains monuments, places and workspaces, the red ovals show scientific from measuring split, the orange ovals contain art split from works, respectively.

Detailed drift recordings show that, e.g. at 40x24 resolution, between 1796-1845, the drift rate, represented as a green line, is 19-22%, whereas between 1960-2009, it is 15-27.5% (Fig. 5-11 and Fig. 5-12). The corresponding lists identify at-risk terminology in the specified periods. Both at-risk terminology lists and the drift diagrams are available for every indexing level, i.e. in resolutions 20x12, 40x24, 50x30 and 60x40, and more.


**Fig. 5-11**. Fluctuating level 2 term drift, term split and term merge rates between 1796-1845.


**Fig. 5-12**. Fluctuating level 2 term drift, term split and term merge rates between 1960-2009.

Finally, our working hypothesis about vector fields generated by ESOM as a suitable model of Trier's semantic fields was confirmed by the analysis of the different term-document (feature-artefact) matrices on indexing level 2. Both in the first and the second period, over ten epochs each, we found index terms with a related meaning clustered onto the same grid nodes (Fig. 5-13 and Fig. 5-14). Examples include e.g. [*mammals, animals*]; [*testament, bible*]; [*sex, relationships*]; [*punishment, crime*]; [*myths, gods, classical, heroes*]; [*scientific, measuring*]; [*machinery, crafts, tools*]; [*plants, flowers*]; [*death, birth*], and many more. These semantic fields were confirmed by HCA as well.

On the other hand, lvl3 results suffered from term volatility and the composition of the actual sample, ultimately due to decisions about what to collect and/or archive. However, we also observed

semantically unrelated terms mapped to the same BMU, i.e. into the same cluster, which raises the possibility that these express syntactically related concepts, i.e. topics. The investigation of this, and how such semantic composites relate to topic shifts, will require future work.

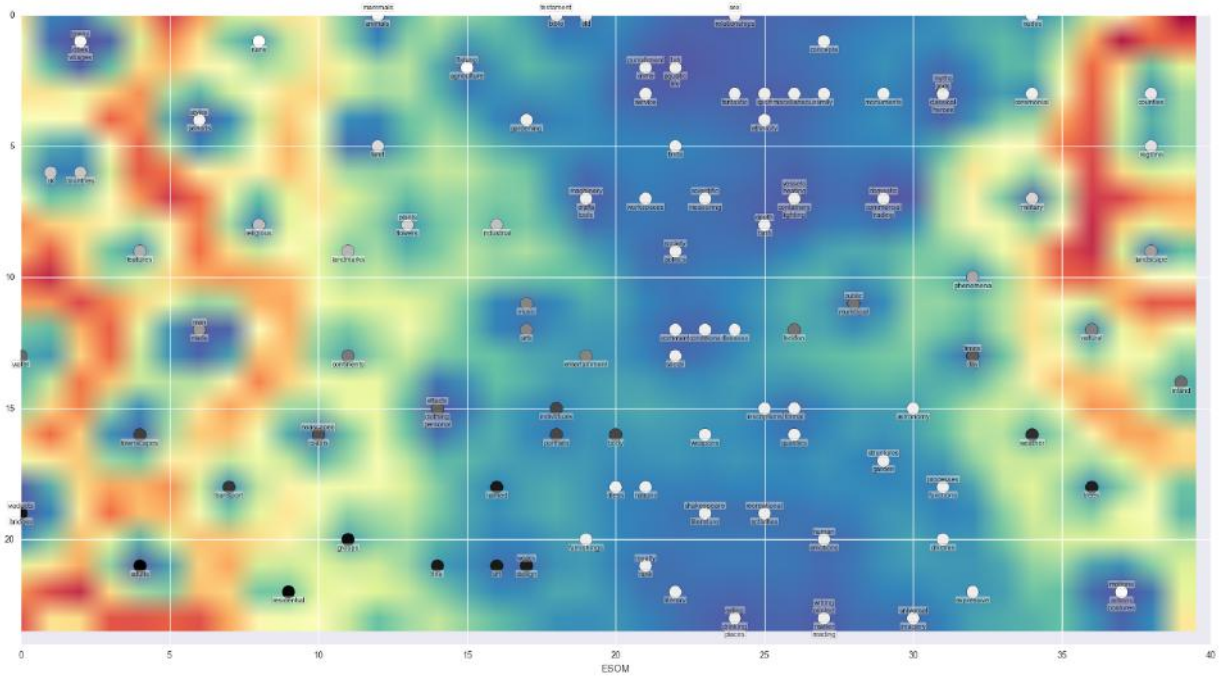

**Fig. 5-13**. Level 2 term clusters as semantic fields in 1796-1800 (40x24 resolution).

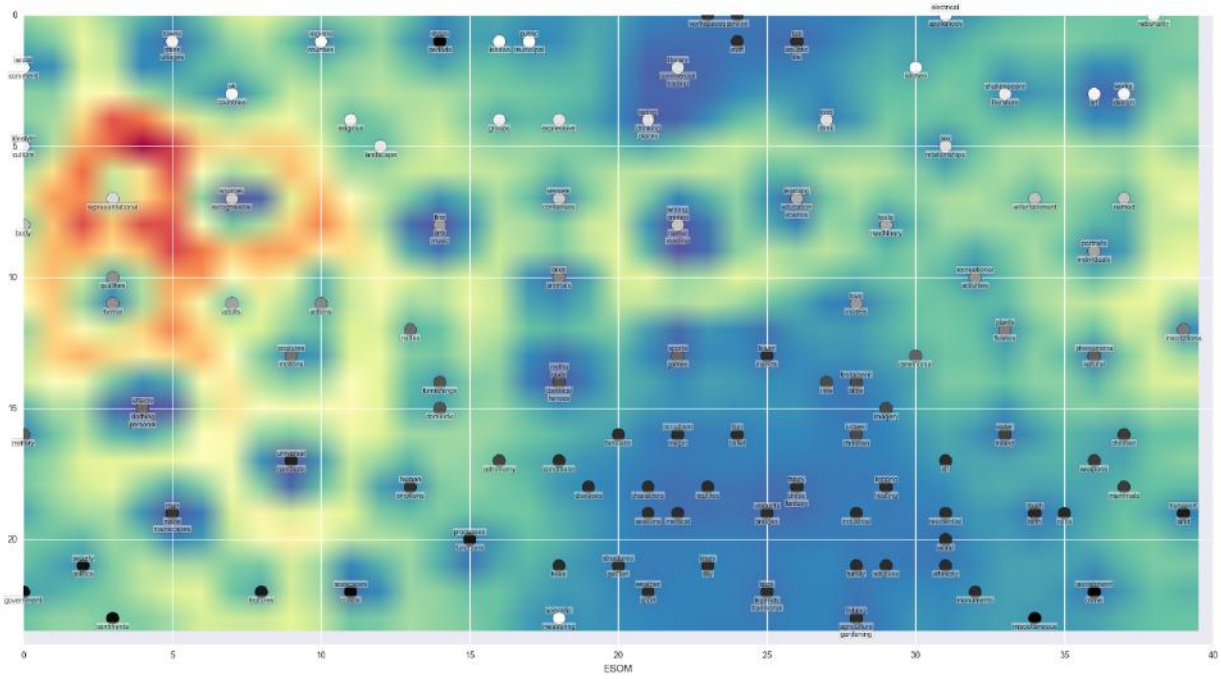

**Fig. 5-14**. Level 2 term clusters as semantic fields in 1960-1964 (40x24 resolution).

### *Future Research*

Our above findings are being communicated to different relevant scientific communities. Further, they feed into implementations in T3.5.4 and T5.3.3 as follows:

- Based on content structure, drift logs make a thresholded alert system possible. This is suitable for DP as a risk management component for the observation of at-risk-terminology, and at the same time helps quality assurance (QA) for collection diagnostics.
- Based on tension structure, with more frequent sampling of real time input, the evolving semantics of a collection can be modelled as a vector field in change. The more accurate such a model, the better predictions it will be able to give for expected changes in the future.
- If we replace index terms (i.e. word semantics) by RDF statements, and tag object content by short statements (sentences) of <Subject-Verb-Object> structure, the field model can be extended to handle sentence semantics as well. Thereby the LRM and the field model can be merged and a next curation tool designed, one that combines ontology-based indexing with vector field semantics. Such a new experiment in the arts domain is in progress.
    - For every system state $A_{tq}$ there exists a respective ontology state $O_{tq}$ that describes the content classes and their logical etc. relationships. Clearly, with automation resulting in scalable systems of content, ontology maintenance becomes a cooperation problem between multivariate statistics and logic. To this end, vector space representations of phrase and sentence content do exist [Padó & Lapata, 2007; Baroni & Lenci, 2010; Socher et al., 2012; Blacoe et al., 2013; Grefenstette et al., 2013; Widdows & Cohen, 2016], i.e. to display e.g. the evolution of the RDF composition of ontologies in an X-ray like manner is entirely possible. The workflow could be e.g. that due to change in the ontology, the modified dependencies cause a top-down trickle-down effect with cost implications as a function to graph modification and maintenance. Vice versa, statistics-induced changes in the ontology graph may trigger a bottom-up value propagation scenario, with similar cost implications.
- The software suite underlying our semantic drift observatory can provide high-quality input for ontology maintenance, pertinent both to WP3 and WP5[57].

## 5.3.2. Semantic Change through Ontology Evolution

**Ontology evolution** can be defined as the process of an ontology change in size and management to accommodate dynamic changes and knowledge interchange in industrial and academic applications. This phenomenon mandates the need for an efficient monitoring and management process [Stojanovic et al., 2002]. In PERICLES, evolving semantics are monitored through methodologies in the field of semantic and concept drift as the vehicle to measure and manage change. A thorough review of meanings and methodologies in this field was presented in Section 5.1, disambiguating the terms *semantic drift*, *concept drift* and *semantic change*.

This section presents an applied methodology in PERICLES stemming from previous work in concept drift, i.e. a change in the meaning of a concept over time, location, culture etc. In these previous studies [Wang et al., 2011], highly applicable notions and metrics for measuring concept drift in the context of data mining, have successfully been transferred to semantic drift.

In detail, the method to measure concept drift in semantics considers two basic pillars of change: (a) the different aspects of change, and (b) whether concept identity is known or not. The different types of change, reflecting its meaning, include:

- **Label**, which refers to the description of a concept, via its name or title;
- **Intension**, which refers to the characteristics implied by it, via its properties;
- **Extension**, which refers to the set of things it extends to, via its number of instances.

[57] The developed software algorithms are available at https://github.com/MKLab-ITI/pericles-semantic-drift, while all relevant datasets can be found at at: https://www.dropbox.com/sh/sbs2nvkjxxjezy7/AABoT0SWEvMlWG7BKdKgbp-8a?dl=0.

Meanwhile, the correspondence of a concept across versions can be either known or unknown, resulting in two different approaches for measuring change:

- **Identity-based** approach (i.e. known concept identity): Assessing the extent of shift or stability of a concept's meaning is performed under the assumption that a concept's identity is known across ontologies: ontology A, and its evolution, ontology B. In other words, each concept of A is considered to correspond to a single, known concept of B.
- **Morphing-based** approach (i.e. unknown concept identity): Each concept is pertaining to just a single moment in time (ontology), while its identity is unknown across versions (ontologies), as it constantly evolves/morphs into new, even highly similar, concepts. Therefore, its change has to be measured in comparison to every concept of an evolved ontology.

The contribution of PERICLES in this area is the adoption, implementation and extension of these methods, in an open, reusable software solution, which is so far lacking. The rest of the section describes our proposed method to measure drift, the dataset synthesized for a proof-of-concept scenario and its results. Future work presented in the end of Section 5.3.2 promises to mend many shortcomings in the field of semantic change by providing an open, domain-independent toolbox.

### *METHOD DESCRIPTION*

Our proposed method adopts the **morphing-based approach**. This choice has been made since this approach is more abstract and generally applicable, as it does not require user input (via an interface, additional concept annotations or explicit concept identities) to pinpoint the correspondence of concepts across ontology versions. On the contrary, the morphing-based method requires as input only a set of ontology versions, **ordered according to the course of change** e.g. time or locations.

As output, for each concept in each version, the method generates three measurements of change (label, intensional and extensional) against concepts in the next ontology in order. For each version, it also generates the average concept change to the next version, for all concepts and for each of the three types, presenting an overview of concept change or stability across versions.

In detail, in order to measure change, the meaning of each concept at a given point $t$ (e.g. in time) is defined as a set of the three different aspects, as follows:

$$C^t = < label_t(C), int_t(C), ext_t(C) >$$

where $C^t$ denotes the meaning of concept $C$ at point $t$, $label_t(C)$ denotes the label aspect of concept $C$ at point $t$, $int_t(C)$ denotes the intensional aspect of concept $C$ at point $t$ and $ext_t(C)$ denotes the extensional aspect of concept $C$ at point $t$.

Furthermore, each aspect can be measured as follows:

$$label_t(C) = o, o\ in\ (C, rdfs{:}label, o)$$

$$int_t(C) = \{\ all\ triplets\ (C,p,o)\ U\ all\ triplets\ (s,p,C)\ |\ p\ is\ owl{:}ObjectProperty\ or\ owl{:}DatatypeProperty\ \}$$

$$ext_t(C) = \{\ i\ |\ i\ in\ (i, rdf{:}type, C)\ \}$$

In other words, the label aspect is given by the `rdfs:label` of a concept. The intensional aspect is a set comprised of the union of all RDF triples with $C$ in the subject or object position of OWL Object Properties or OWL Datatype Properties. The extension aspect is defined as the set of all instances of `rdf:type` $C$. Overall, label is a string, intension is a set of triples and extension is a set of strings.

Based on these definitions and using appropriate similarity metrics one can measure the change/evolution of aspects across versions of the ontology. Table 5-3 summarises the metrics adopted in our approach, building upon the state-of-the-art. In all cases, each concept of an ontology version is compared to all concepts of the version next in order, for each of the aspects.

**Table 5-3**. Similarity metrics to measure concept differentiation across ontology versions in PERICLES.

| Aspect | Similarity Metric to Measure Differentiation |
|---|---|
| Label | String similarity with Monge-Elkan |
| Intention | Jaccard similarity between sets of triples |
| Extension | Jaccard similarity between sets of strings |

For **label drift**, the two strings are compared based on the text similarity algorithm Monge-Elkan [Monge & Elkan, 1996], which empirically shows the best results for strings found in ontologies, such as CamelCase and snake_case, without having to trim or split them. More precisely, in the morphing-based approach, the *label drift* of a concept $C$ between versions $t_1$ and $t_2$, is defined as the average of Monge-Elkan text similarity between $C$ in $t_1$and all concepts of $t_2$.

$$label_{t_1 \to t_2}(C) = \frac{\sum_{i=1}^{n} MongeElkan\left(label_{t_1}(C), label_{t_2}(C_i)\right)}{n}$$

In order to measure the similarity of two sets, we deploy the Jaccard similarity [Jaccard, 1902; Jaccard, 1912], which is defined as follows:

$$Jaccard(A,B) = \frac{A \cap B}{A \cup B}$$

where $A, B$ are two sets of items. Based on that, we define the **intensional drift** of a concept $C$ between versions $t_1$ and $t_2$ as the average of the Jaccard similarities between $C$ in $t_1$and all concepts of $t_2$. This is defined as:

$$int_{t_1 \to t_2}(C) = \frac{\sum_{i=1}^{n} Jaccard\left(int_{t_1}(C), int_{t_2}(C_i)\right)}{n}$$

where $int_{t_1 \to t_2}(C)$ is the intensional drift of $C$ between versions $t_1$and $t_2$, $int_{t_1}(C)$ is a set of triples representing the intension of $C$ at point $t_1$ (properties) and $n$ is the total number of concepts in $t_2$.

Similarly, we define the **extensional drift** of concept $C$ between versions $t_1$ and $t_2$ as the average of the Jaccard similarities between $C$ in $t_1$and all concepts of $t_2$.

$$ext_{t_1 \to t_2}(C) = \frac{\sum_{i=1}^{n} Jaccard\left(ext_{t_1}(C), ext_{t_2}(C_i)\right)}{n}$$

where $ext_{t_1 \to t_2}(C)$ is the extensional drift of $C$ between versions $t_1$and $t_2$, $ext_{t_1}(C)$ is a set of strings representing the extension of $C$ at point $t_1$ (instances) and $n$ is the total number of concepts in $t_2$.

Finally, the **total drift** of concept $C$ between versions $t_1$ and $t_2$, is defined as the average of label, intensional and extensional drift between the same versions:

$$whole_{t_1 \to t_2}(C) = \frac{label_{t_1 \to t_2}(C) + int_{t_1 \to t_2}(C) + ext_{t_1 \to t_2}(C)}{3}$$

### *IMPLEMENTATION*

The above method was implemented as a software tool, in order to reproduce the results and apply the methods in multiple occasions, not only within the framework of PERICLES, but also beyond the DP field, encouraging domain-independent semantic drift research, while disseminating the project's results. The current version of the software tool is implemented as a command-line cross-platform

application, using Java. The OWL-API library[58] was used to handle RDF/OWL operations, while the Simmetrics library[59] provided the implementation of Monge-Elkan text similarity measure. The implementation part shows much room for improvement and extension in future work, listed in the end of the section, promising to contribute significantly to the Semantic Web and semantic drift areas.

### *Dataset*

In order to validate the approach and apply the methodology in the PERICLES domain, a realistic dataset was synthesized, by extending the SBA ontology. The dataset is comprised of SBA ontology versions across time, modelling the evolution and drift of three concepts in the A&M domain, namely Computer-based (CB), Mixed-Media (MM) and Software-based (SB) concepts. The dataset may be synthetic, but is still realistic, as it was based on an internal Tate report that describes the changes to Tate's cataloguing of 8 SBA artworks in the period 2003-2013. Thus, the dataset contains a total of 9 semantic models for this period, one model per year, excluding the years when no changes in the cataloguing occurred.

### *Results*

The proof-of-concept use case to measure semantic drift in the A&M domain was performed by feeding the extended SBA ontology versions, ordered by year, to the software tool. The output is presented here, starting from morphing chains for each of the concepts, showing their interrelations in-depth, then an overall graph showing the different measures of stability across versions and finally a concept stability matrix. Concept stability is measured as similarity, in the range of 0, for completely disjoint, to 1, for entirely identical label/properties/instances (according to label/intensional/extensional drift respectively).

Initially, morphing chains show in detail concept similarity for each aspect, and how concept meanings migrate from one concept to another, each year from 2003 to 2013. Inspecting the label aspect shown in Fig. 5-15, it is apparent that the highest similarity measure holds between concepts with the same name across versions, demonstrating stability.

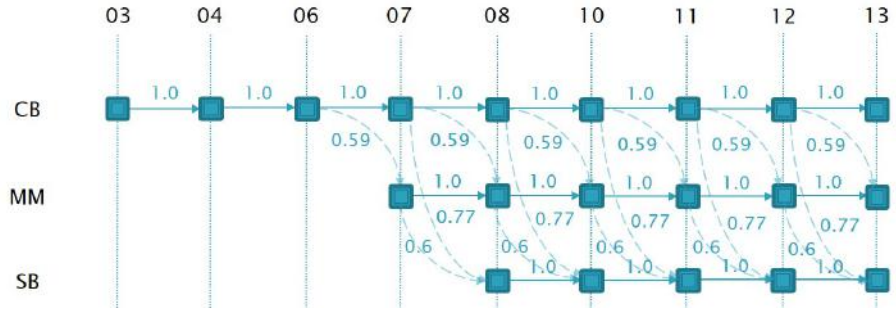


**Fig. 5-15**. Morphing chains for the label aspect.

Likewise, the intensional aspect, shown in Fig. 5-16, demonstrates equal stability, as properties do not vary significantly across versions.

[58] http://owlapi.sourceforge.net/
[59] https://github.com/Simmetrics/simmetrics

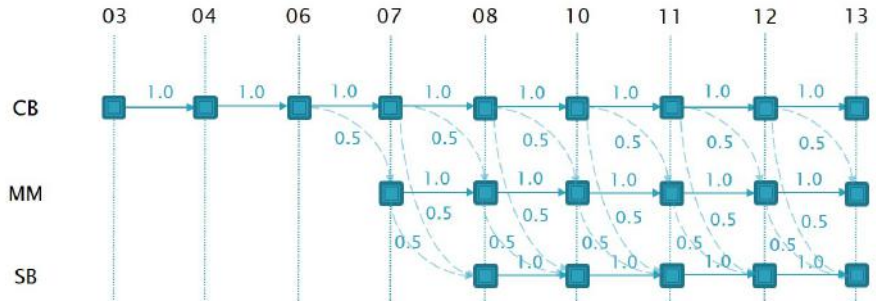


**Fig. 5-16**. Morphing chains for the intensional aspect.

On the contrary, the extensional aspect, shown in Fig. 5-17, demonstrates variations from version to version, with the most significant ones being the complete migration of the CB media concept partially to MM and to SB concepts, due to its instances shifting type.

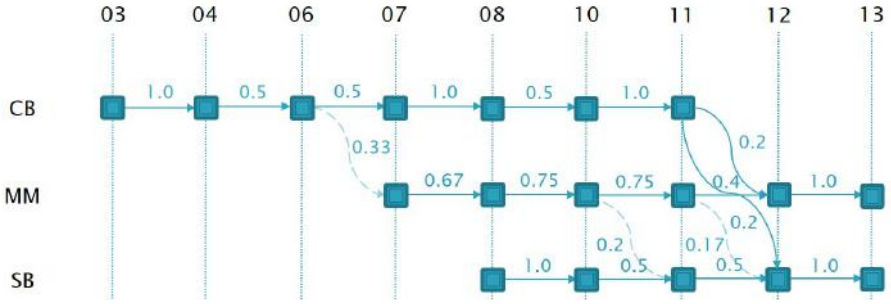


**Fig. 5-17**. Morphing chains for the extensional aspect.

Finally, the whole aspect depicts these changes in the greater scale, reflecting stabilities (due to label and intension) and some instabilities (due to extension).

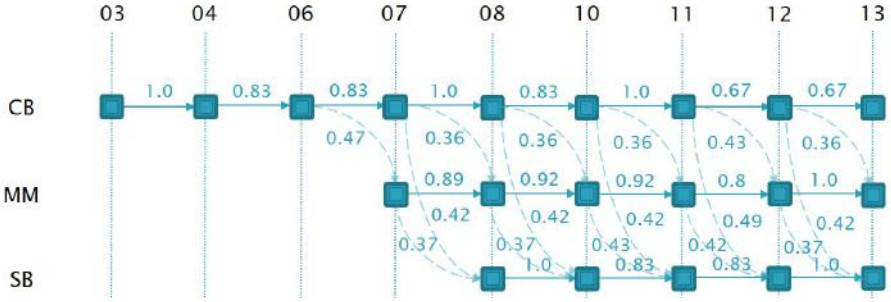


**Fig. 5-18**. Morphing chains for the whole aspect.

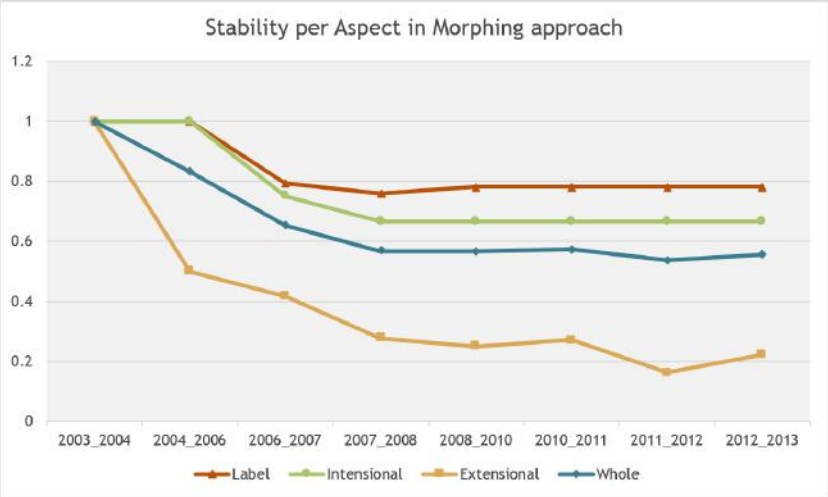


**Fig. 5-19**. Concept stability over time for the different aspects.

Averaging each aspect for all concepts per version reveals concept stability over time as shown in Fig. 5-19. This revealing graphic representation clearly shows at a glance that:

- The label aspect is the most stable aspect, followed by intension, since labels and properties remain quite constant in the sample dataset.
- The extensional aspect is the least stable, as all instances of CB type are eventually evolved into MM or SB.
- Stability is reduced in all aspects during 2003-2008, as the ontology is enriched with new concepts.

**Table 5-4**. Overall stability and ranking per concept and per aspect across all versions.

| **Label** | | |
|---|---|---|
| **Rank** | **Concept** | **Stability** |
| 1 | CB | 0.826 |
| 2 | SB | 0.799 |
| 3 | MM | 0.736 |

| **Extensional** | | |
|---|---|---|
| **Rank** | **Concept** | **Stability** |
| 1 | CB | 0.270 |
| 2 | MM | 0.268 |
| 3 | SB | 0.253 |

| **Intensional** | | |
|---|---|---|
| **Rank** | **Concept** | **Stability** |
| 1 | CB | 0.722 |
| 2 | MM | 0.667 |
| 3 | SB | 0.654 |

| **Whole** | | |
|---|---|---|
| **Rank** | **Concept** | **Stability** |
| 1 | CB | 0.606 |
| 2 | SB | 0.568 |
| 3 | MM | 0.557 |

Averaging each concept per aspect for all versions shows a measure of overall stability, as listed in Table 5-4. The CB concept appears to be the most stable one in all aspects. While this concept ranks first, examining the model more closely reveals that this could actually be attributed to its high similarity to the other two concepts. However, this fact cannot be attributed to stability per se. Due to the morphing-based approach comparing to all concepts and this particular concept being highly similar to the other few concepts, similarity across concepts can be falsely perceived as stability here, revealing a limitation of the method. In other words, high similarity across concepts can be interpreted as stability. This limitation is not necessarily misleading, as it could be inherently lifted when enriching the synthetic ontology with more concepts. Meanwhile, this is also a pertinent

feature to the morphing-based method, dominated by uncertainty, while in an identity-based method the issue would disappear.

#### *FUTURE WORK*

Future research directions aim to broaden the scope of domain-independent, open tools enabling the Semantic Web community and disseminating the project's results. So far the core methods developed in PERICLES for calculating drift measures based on ontology evolution have focused on the morphing-based approach, due to its generality and its low requirements from the user (i.e. the ordered set of ontology versions). On the other hand, an identity-based method should also be implemented in the future, as it entails far less uncertainty, giving a much clearer picture of concept stability and drift insights. However, it requires user input to indicate the correspondence of a concept across versions, either through metadata or a GUI for user interactions.

The methods themselves can always be enriched with more efficient similarity metrics as done in the current morphing-based methods. As metrics vary, new insights may emerge stemming from limitations. E.g. some metrics for stability may further require normalization.

Furthermore, there are many improvements to implement for both approaches. While the core morphing method is complete, it should be accompanied by a GUI to input basic values such as to indicate file input, order and obtain results graphically, such as those presented in this section. Meanwhile, after implementing the identity approach as well, a GUI will be an even greater facilitator, allowing the user to connect corresponding concepts across versions using graphical means. The tools are planned to be implemented as both standalone cross-platform applications (using JavaFx), or even as Protégé plugins. The latter being a very popular and versatile platform in the community will greatly accelerate adoption and dissemination efforts.

#### *DISCUSSION*

This subsection presented a method for measuring semantic change in terms of semantic concept drift. Methods in state-of-the-art have been optimized and adapted, measuring label, intensional, extensional and whole (total) drift, inspired by methods in the field of Machine Learning, and following the generic, morphing-based approach. The method has been implemented as a domain-independent, cross-platform software tool that will help stimulate research in the area and disseminate the project's results. Consequently, a proof-of-concept experimentation has been performed, by synthesizing a realistic dataset from A&M reports from Tate, showing concept drift in terms of morphing chains, aspect measures and concept stability across time (from 2003 to 2013). The tool shows promise to be extended with more methods and a GUI to facilitate adoption[60].

Regarding limitations, an issue arises when considering the concept stability measure shown in Table 5-4. It seems that the CB concept is the most stable one, but actually it ranks first because of its high similarity to the other two concepts. This is a feature pertinent to the morphing-based method, dominated by uncertainty and lifted as the ontology grows in size, beyond a synthetic dataset. Notably, it would also be extinct when using an identity approach at the cost of manual labor to annotate the corresponding concepts.

### *5.3.3. Studying Community Change*

In the following set of experimentation, we report investigations into the potential users of media data. The OAIS model refers to a '**designated community**' as a way of understanding potential users or consumers of a particular set of information [Vardigan & Whiteman, 2007]. This concept is used in

[60] The software and all relevant datasets are available at: https://www.dropbox.com/s/f150rxbx2bi7lvj/SemanticDriftMetrics.zip?dl=0

the development of an OAIS-compliant information system, in order to elicit requirements for that information system. Within the preservation context, this may be likened to the development of 'personas' in the user design of technology. A 'designated community' is certainly a relatively abstract concept, and hence problematic as a design artefact: it is noticeable that in practice the concept is often used to refer to characteristics common to large numbers of individuals within the 'designated community' demographics (i.e. language proficiencies, interests, educational roles or tasks).

**User personas** are often applied in order to develop a more concrete scenario for technology use (cf. [Pruitt & Grudin, 2003; Community Systems Group, 2007]). The intent of development of a user persona is to provide a pragmatically usable 'target' for a development process, so that designers can consider the impact of design decisions on specific users, rather than considering the problem in the abstract. It is our contention here that 'designated communities' function in a somewhat similar manner. Indeed, as Allinson notes in [Allinson, 2006], designated communities may be viewed as being formed from a complex set of individuals, each of which contribute to this user community.

Therefore in the following analysis, we focus on describing the kinds and types of users which contribute to 'community'. Although, community can be studied using practical measures such as using 'foot fall' metrics (e.g., by monitoring visitor behaviour) and surveys, we note that this has limited value in studying the larger community of users. Therefore, by systematically considering the interactions and functions between art objects and users, we are able to consider the processes involved in order to gain a greater idea of the user community. In particular, we note that (a) the object exists within a gallery's collection; (b) a catalogue provides an interface between the object and potential users; (c) users wishing to access the object (sometimes termed 'community').

In this model, we note that there are two ways of understanding users: firstly, by studying catalogue information (b), and secondly, by studying (c) potential and/or existing users. These two approaches therefore give different perspectives of the same phenomenon, as follows.

The catalogue information explored in the first approach operates as an interface between object and user. However, the catalogue is not a passive bystander in this process, since it is actively created (curated) by an archivist who is aware of the cultural context of the usage of the art objects, and thus we can regard this interface in some ways also as a 'filter' (we note that in turn the art objects themselves are a product of the artist's awareness of, and interaction with, their cultural context). Through analysing this catalogue data we can then identify potential differences in art object usage over time or across art institutions. It is important to keep in mind that object appraisal and cataloguing can be understood and modelled as situated actions [Iivari & Linger, 1999], which occur within a specific and varying cultural and financial context, so that several factors are involved in shaping practices within a given time and space.

In the second instance, by studying users who may wish to, or already, access the art object we are better able to understand who these users might be and their interests and requirements. As already noted, this is unlikely to be one specific purpose, but most likely a range and variety of complementary and potentially conflicting needs. In order to do this, we refer specifically to data collected via social media which captures individual's own documentation of their interactions with an art institution. By analysis of this social media data, we are able to better understand the range of uses (and types of users) interacting with the institution.

We note that this approach - used to identify different users - may be applied both over time and within a specific time period, potentially comparing across institutions; we believe that both of these aspects of different variety in users is valuable for providing a holistic view of the contextual ecosystem in which art object, art institutions, and associated users operate.

### *Experiment 1: Examining the Cultural Interface to Users*

In this experiment, we examine the adaptation of community change through the interface of the art museum catalogue, specifically the title which mediates between the object itself and the user. Here we initially focus on data from four art museums for the past 1016 years (year 1000 to date), and on the basis of these results focus on data from 1900 to date; we do this in order to study in more detail one particular institution (Tate) in comparison to a wider range of art museums.

**Data Collection and Analysis**

We use a database assembled from publicly available art catalogues relating to four large English-language museums. These are: the Getty (268,080 items), the Metropolitan Museum (424,065), MOMA (137,381), and the Tate galleries (69,201). Collection of these resources was conducted using publically available datasets from these museums: the latter two institutions published their collection catalogue as .csv or .json files; for the former two institutions, this was web scraped via their online catalogue (using a Python script based around the BeautifulSoup Python HTML parser and the Mechanize Python browser module). Before inclusion, the different datasets were cleansed and basic information (e.g. title, data of creations, date of acquisition, artist nationality) was converted into a consistent format. To examine the interface between art object and perceived audience, we analyse the language description assigned to art objects via their titles.

In our analysis, we first consider data relating to art objects that contain a title text and have a creation date between the years 1000 and 2016. This gives 186,848 cases in our data set, and totals around 1 million words (1,030,978 words; mean = 5.5 words pers title). Like in the semantic drift experimentation reported in Section 5.3.1, we adopt the UCREL semantic tagger (via Wmatrix tool [Rayson, 2008]), and use this to generate semantic information for the art object title text.

In our analysis of the title data, we are interested in how the titles represent themes, in how the art objects are portrayed, and how these change over time, potentially indicating adaptation to the user community and cultural context.

To examine this change, we performed clustering of the art object titles by century, using hierarchical clustering and more specifically the complete linkage method implemented in R[61]; the latter makes use of maximal distance between components in a cluster to define clustering distance. All columns that sum to zero were removed, and all NAs/empty cells were replaced by zeroes.

**Semantic Temporal Clustering Results**

Fig. 5-20 demonstrates the way in which the themes from art object titles cluster over the past 1000 years. In particular, we can see that there is a main division between artworks before and after 1500; after this date, there are main groupings by centuries consisting of 1500s and 1600s, the 1700s and 1800s, and finally, the 1900s and 2000s.

[61] As described in http://www.r-tutor.com/gpu-computing/clustering/hierarchical-cluster-analysis

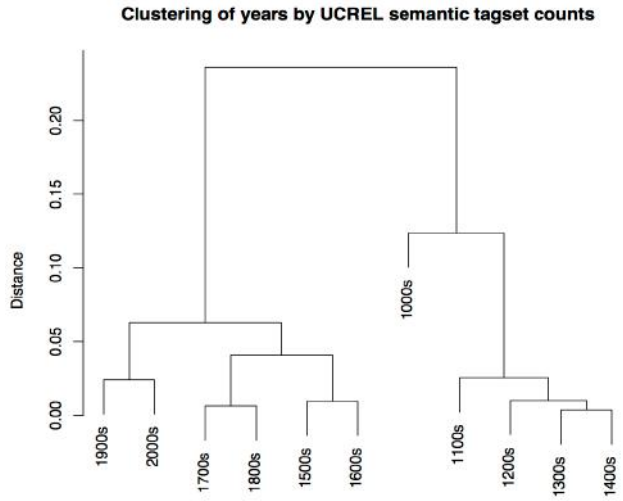


**Fig. 5-20**. Dendrogram showing Semantic Clustering of Centuries by UCREL Semantic Tag usage; note clear clustering of modern, 1500s-1800s and prior centuries.

**Semantic Tag Content Analysis**

Having identified that semantic tags of titles for the most recent artworks across the four museums in the data set cluster together (the 1900s and 2000s), we now focus on how analysis of the specific time period can inform us about the relevant user communities.

To do this, we perform corpus comparison for a sub-sample (title data relating to Tate) against the larger title data set relating to the four museums. These two data sets are as follows (note that Tate data is also included in the four museums data, and is retained for this comparison): Number of cases is 17,242 (Tate), 128,232 (four museums); Total number of words, 60,397 (Tate), 703,690 (four museums); Average words per title, 3.5 (Tate), 5.5 (four museums).

Statistical comparison of the frequency of semantic tag terms was conducted in Wmatrix using the log-likelihood measure; here 'overuse' ('+' or '-') is reported to represent characteristic features of the respective art museum (Tate or the four museums). These results are reported in Table 5-5 and Table 5-6.

Table 5-6 shows the top 50 overused/underused items resulting from the analysis. In Table 5-5 we discuss these in more detail for the subset of the Tate. All of these findings have a statistical significance of greater than $p<0.001$ (the log-likelihood value can be interpreted using the chi-squared distribution and respective critical values).

Looking in more detail at the characteristics of the Tate semantic categories (i.e., those which are used proportionately more in this data; Table 5-5), we can see that these can be grouped into the following main themes: humans (People: Female, Anatomy and physiology, Kin, People, People: Male), nature (Living creatures: animals, birds, etc., Plants, Farming & Horticulture) and the outdoors (Geographical terms, Sailing, swimming, etc., Weather). In addition, we notice semantic categories relating to orientation or format, which may relate to the content or the format of the art object (Colour and colour patterns, Arts and crafts, Shape, Stationary, Location and direction, Putting, pulling, pushing, transporting, Damaging and destroying). We also note that Geographical names and Electricity and electrical equipment are individual semantic categories, but may be relatively specific to the Tate collection (especially the Geographical names, which amongst others include London).

**Table 5-5**. Top 25 Semantic tag categories overused in Tate catalogue titles (RF = relative frequency of usage in the Tate data).

| Semantic category | Tag ID | RF | Top title items |
|---|---|---|---|
| Colour and colour patterns | O4.3 | 1.92 | red, black, blue, yellow, green |
| Time: Period | T1.3 | 1.42 | night, morning, day, summer, evening |
| Living creatures: animals, birds, etc. | L2 | 1.27 | dog, elephant, horse, sheep, bird |
| Plants | L3 | 0.91 | tree, garden, trees, flowers, flower |
| People: Female | S2.1 | 0.9 | woman, girl, women, girls, female |
| Arts and crafts | C1 | 2.07 | drawing, painting, sculpture, self-portrait, art |
| Shape | O4.4 | 0.77 | square, line, spiral, lines, vertical, cross |
| Geographical terms | W3 | 1.21 | landscape, sea, earth, mountain, river |
| Anatomy and physiology | B1 | 1.41 | head, skull, hand, eye, torso |
| Kin | S4 | 0.54 | mother, family, wife, parents, son |
| Stationary | M8 | 0.18 | seated, sitting, still, stays, settled |
| Sailing, swimming, etc. | M4 | 0.34 | harbour, port, bather, boats, boat |
| Location and direction | M6 | 1.21 | reclining, interior, standing, this, centre |
| Geographical names | Z2 | 5.61 | Lebanon, Saida, London, studio_Shehrazade, south_Lebanon |
| Putting, pulling, pushing, transporting | M2 | 0.4 | seated, dropping, hanging, suspended, set |
| People | S2 | 0.37 | child, children, people, human, person |
| Weather | W4 | 0.23 | Snow, wind, rain, storm, mist |
| People: Male | S2.2 | 0.5 | man, boy, men, male, boys |
| Electricity and electrical equipment | O3 | 0.16 | circuits, electric, robot, battery, plug |
| Damaging and destroying | A1.1.2 | 0.18 | broken, fragment, destruction, crash, ruins |
| Sensory: Sound | X3.2 | 0.11 | sound, listening, heard, noises, siren |
| Farming & Horticulture | F4 | 0.27 | field, fields, farm, peasant, landscape |
| Participating | S1.1.3+ | 0.1 | collaboration, meeting, forum, reunion, bacchanal |
| Without clothes | B5- | 0.22 | nude, naked, stripped, nudes, topless |
| Infrequent | N6- | 0.04 | once, rare, occasional, occasionally |

Interpreting these results as a representation of the collection in terms of user community context, this shows that there is a greater concern for the coverage of humans, nature and the outdoors, as well as specific geographic locations, relative to the overall collection of the four museums. Here we see that there is a focus on topics which are relevant to the Tate user community over the past century: the high esteem with which figurative work is held, as well as a sense of (or desire for) connection to the natural world and landscape; there are also the location specific details, indicating London, as well as more distant, relevant places. As well as the obvious potential users of the galleries, such as art scholars and members of the public, it is also worth considering here the role of other users for whom these titles were relevant: for example, stakeholders such as funding bodies and governmental organisations can also be considered relevant to the user community, and so may have influenced the content of a particular gallery (although we expect this to be more directly a result of collection policy and thus more likely to be observable over a shorter time period, such as at the decade level).

Since these titles provide an insight into the interface between art object and user, we can propose that there is a possible change/evolution in the way that art objects are presented to users, both over time across the four museums, and also in the most recent stable groups of semantics of art object titles (for the Tate versus the four museums). To some extent this can be regarded as the cultural context of these art objects and more widely of the art institution. However, what this analysis cannot disentangle is a distinction between the adaptation in the way an art object is presented to the user by the curator/archivist as a result of the community, and how an artist might differently represent his/her work as a result of that artist's perception of the intended user community, i.e.preconceptions regarding reception of the work. Similarly, it is not clear how this process might be affected by gallery funding requirements or collection policy; however, we can assume that these all respond in some way to requirements of some form of 'user' and/or 'community'. We explore this further in the following experiment.

**Table 5-6**. 50 most significant title semantic tags for Tate subset versus four museums (O1 = observed frequencies in Tate data, RF1 = relative frequencies in Tate data; O2 = observed frequencies in four museums data; FR2 = relative frequencies in four museums data; LL = Log-Likelihood statistic; "+" = overuse by Tate, "-" = overuse in four museums data).

| Semantic category | Tag ID | O1 | RF1 | O2 | RF2 | +/- | LL |
|---|---|---|---|---|---|---|---|
| Paper documents and writing | Q1.2 | 381 | 0.7 | 22702 | 4.09 | - | 2221.32 |
| Objects generally | O2 | 1117 | 2.04 | 26637 | 4.8 | - | 1027.69 |
| Colour and colour patterns | O4.3 | 1051 | 1.92 | 3656 | 0.66 | + | 756.65 |
| Quantities | N5 | 571 | 1.04 | 2338 | 0.42 | + | 312.04 |
| Time: Period | T1.3 | 774 | 1.42 | 3961 | 0.71 | + | 259.59 |
| Living creatures: animals, birds, etc. | L2 | 692 | 1.27 | 3720 | 0.67 | + | 203.32 |
| Plants | L3 | 496 | 0.91 | 2348 | 0.42 | + | 200.97 |
| Numbers | N1 | 3253 | 5.95 | 42308 | 7.62 | - | 199.58 |
| People: Female | S2.1 | 490 | 0.9 | 2392 | 0.43 | + | 184.78 |
| Arts and crafts | C1 | 1132 | 2.07 | 7332 | 1.32 | + | 177.18 |

| Semantic category | Tag ID | O1 | RF1 | O2 | RF2 | +/- | LL |
|---|---|---|---|---|---|---|---|
| The Media: Books | Q4.1 | 110 | 0.2 | 3182 | 0.57 | - | 164.54 |
| Shape | O4.4 | 421 | 0.77 | 2078 | 0.37 | + | 154.58 |
| Geographical terms | W3 | 663 | 1.21 | 3898 | 0.7 | + | 148.27 |
| Measurement: Length & height | N3.7 | 18 | 0.03 | 1086 | 0.2 | - | 107.03 |
| Anatomy and physiology | B1 | 771 | 1.41 | 5530 | 1 | + | 74.71 |
| Kin | S4 | 293 | 0.54 | 1672 | 0.3 | + | 72.14 |
| Comparing: Similar | A6.1+ | 31 | 0.06 | 1098 | 0.2 | - | 71.84 |
| Stationary | M8 | 100 | 0.18 | 352 | 0.06 | + | 70.69 |
| Sailing, swimming, etc. | M4 | 188 | 0.34 | 922 | 0.17 | + | 70.11 |
| Location and direction | M6 | 659 | 1.21 | 4724 | 0.85 | + | 64.05 |
| Geographical names | Z2 | 3068 | 5.61 | 27205 | 4.9 | + | 48.48 |
| Putting, pulling, pushing, transporting | M2 | 221 | 0.4 | 1330 | 0.24 | + | 45.66 |
| Linear order | N4 | 312 | 0.57 | 4568 | 0.82 | - | 43.69 |
| People | S2 | 202 | 0.37 | 1206 | 0.22 | + | 42.9 |
| Weather | W4 | 125 | 0.23 | 652 | 0.12 | + | 39.88 |
| People: Male | S2.2 | 272 | 0.5 | 1792 | 0.32 | + | 39.64 |
| Electricity and electrical equipment | O3 | 86 | 0.16 | 397 | 0.07 | + | 36.86 |
| Damaging and destroying | A1.1.2 | 100 | 0.18 | 499 | 0.09 | + | 35.75 |
| Being | A3 | 1 | 0 | 214 | 0.04 | - | 32.31 |
| Sensory: Sound | X3.2 | 59 | 0.11 | 242 | 0.04 | + | 32.14 |
| Farming & Horticulture | F4 | 149 | 0.27 | 888 | 0.16 | + | 31.83 |
| Participating | S1.1.3+ | 57 | 0.1 | 238 | 0.04 | + | 30.02 |
| Without clothes | B5- | 121 | 0.22 | 693 | 0.12 | + | 29.45 |
| Infrequent | N6- | 20 | 0.04 | 39 | 0.01 | + | 28.22 |
| Darkness | W2- | 33 | 0.06 | 102 | 0.02 | + | 28.16 |
| Measurement: Distance | N3.3 | 42 | 0.08 | 154 | 0.03 | + | 27.82 |
| Substances and materials: Solid | O1.1 | 331 | 0.61 | 2440 | 0.44 | + | 27.48 |
| Like | E2+++ | 5 | 0.01 | 286 | 0.05 | - | 27.31 |

| Semantic category | Tag ID | O1 | RF1 | O2 | RF2 | +/- | LL |
|---|---|---|---|---|---|---|---|
| Relationship: Intimacy and sex | S3.2 | 85 | 0.16 | 443 | 0.08 | + | 27.18 |
| Concrete/Abstract | A1.6 | 55 | 0.1 | 238 | 0.04 | + | 27.01 |
| Long, tall and wide | N3.7+ | 49 | 0.09 | 203 | 0.04 | + | 26.19 |
| Warfare, defence and the army; weapons | G3 | 200 | 0.37 | 1351 | 0.24 | + | 26.1 |
| Vehicles and transport on land | M3 | 228 | 0.42 | 1591 | 0.29 | + | 25.48 |
| Open; Finding; Showing | A10+ | 65 | 0.12 | 1183 | 0.21 | - | 25.13 |
| Light | W2 | 93 | 0.17 | 522 | 0.09 | + | 24.09 |
| Sad | E4.1- | 58 | 0.11 | 272 | 0.05 | + | 24.01 |
| Language, speech and grammar | Q3 | 98 | 0.18 | 1599 | 0.29 | - | 24 |
| Substances and materials: Liquid | O1.2 | 97 | 0.18 | 557 | 0.1 | + | 23.42 |
| Food | F1 | 282 | 0.52 | 2081 | 0.37 | + | 23.26 |
| Existing | A3+ | 255 | 0.47 | 1851 | 0.33 | + | 23.13 |

### *Experiment 2: Inferring User community from Social Media Data*

We have now identified the kinds of content and themes which might be present in the most recent cluster of artworks (also in particular for Tate). We now examine social media to explore what this can reveal about the user community around art institutions, using in particular the case study of Tate. We first provide a description of the social media ecosystem surrounding the institution, and then review in more detail the processes.

#### Twitter Data Collection

To analyse the structure of the social media user community around Tate, data was harvested from Twitter. Tweets were collected previously for the study of social media content elsewhere in this project (reported in D4.3 [PERICLES D4.3, 2016]) using the following process: Since Twitter provides a rate-limited search interface with a fifteen minute time limitation we note that we are only able to retrieve a proportion of search matches. The search interface is authenticated via OAuth and returns query responses in JSON (which are easily interpretable through compatible libraries such as Python's simplejson[62]). We note that other APIs are available for Twitter, but these large scale paid-for services were not seen as suitable for the current study. The search interface gives various modes for querying: here we apply keywords relating to Tate as a trade-off between search precision and recall.

#### Twitter Data Processing

Our initial filtering of Twitter is minimal (i.e. we use the search term ‘tate’ rather than ‘tate gallery’), even though this also returns terms such as ‘state’ or ‘estate’. This is done to increase recall (the fraction of relevant instances retrieved, see [Jardine & Van Rijsbergen, 1971]) at the earliest stage,

[62] https://pypi.python.org/pypi/simplejson/

which enables the widest range of potentially relevant data to be collected, thereby limiting bias at the earliest stages. However, this increases the filtering task required in processing and cleaning up the data, which is non trivial, and therefore iterative in nature [Abel et al., 2012]. Since our approach is exploratory (i.e. we want to characterise the dataset), we do not concern ourselves with optimising the search strategy for a particular profile. Our use of the substring 'tate' in case-insensitive search system captures hash tagged tweets or tagged posts, as well as mentions of the term itself (as noted previously, we retrieve a proportion of false positive terms. For the purposes of the present evaluation we apply strict filtering rules in order to limit material returned to that containing either the string 'tate' with appropriate word boundaries, or material containing the Tate's hostname. Of the original 222,356 tweets collected between 20-Feb-2015 and 20-Nov-2015, the sample resulting from this filtering consisted of 22,000 tweets.

**Social Network Analysis**

Based in the tweets harvested which were relevant to 'tate', users and their respective followers were identified. Mutual (i.e. two-way) following connections were identified between users in this data set, with these users and their connections retained for further analysis. Social network analysis of the dataset was conducted using Gephi.

Layout of the network graph data was arranged using the *Force Atlas 2* algorithm in Gephi[63], using the *LinLog mode*, preventing overlaps, and with *Gravity* set to 0.01. In the graph diagrams appearing in the following subsections, nodes are labelled with the respective Twitter usernames.

**Twitter and the Tate User Community**

The resulting data set consists of the following attributes: there are 172 nodes (i.e. Twitter users) and 858 edges (reciprocal links between them). We note that this is small subset of the total number of Twitter users who may post content relating to Tate, however we believe that this sample is relevant because it relates to Twitter users who are participating in information flow within this social network. To appear in this dataset requires individuals to both follow and be followed by others in the target dataset; since following an individual on Twitter functionally permits individuals to receive the other's updates this implies that individuals are both consumers and producers of information. Indeed, this level of participation indicates a potentially greater involvement within the user community, and those Twitter users who are more relevant to the Tate user community (rather than purely 'broadcasters'). Table 5-7 shows the top ten Twitter users referencing 'tate', which contains a large number of traditional news media organisations who simply broadcast information. Tate (the Twitter username for Tate Galleries) perhaps unsurprisingly features prominently here (at number 4), and we also note the users artnet and Asamsakti who are not traditional broadcasters (the latter of which is a prominent 'art lover' identified in our subsequent analysis).

**Table 5-7**. Top ten Twitter users referencing 'tate' by number of followers.

| Ranking | Twitter User | No. followers |
|---|---|---|
| #1 | BBCNews | 5,619,480 |
| #2 | guardian | 4,832,159 |
| #3 | BritishVogue | 3,060,552 |
| #4 | Tate | 2,289,884 |

[63] http://www.gephi.org

| Ranking | Twitter User | No. followers |
|---|---|---|
| #5 | Independent | 1,652,465 |
| #6 | 1DFAMILY | 1,245,752 |
| #7 | CP24 | 1,186,306 |
| #8 | artnet | 1,165,472 |
| #9 | Asamsakti | 1,099,810 |
| #10 | AJENews | 966,077 |

Descriptive statistics of this network are as follows: average degree=9.977, network diameter=7, and average path length=2.902. This indicates a reasonably high average connectivity (although as we shall see, this is in fact the result of a small number of highly connected users). The network diameter indicates that the furthest nodes may be bridged within seven steps; the average path length is reasonably short, suggesting a fairly closely connected graph more comparable to 'friend' social networks (i.e. a set of Facebook friends) than randomly-chosen individuals (Milgram [Milgram, 1967] famously found an average path length between Americans of 6 steps, causing him to describe the result as 'six degrees of separation'). The resulting graph illustrating the network data is presented in Fig. 5-21. We will return to discuss this after describing specific groupings within this network in more detail.

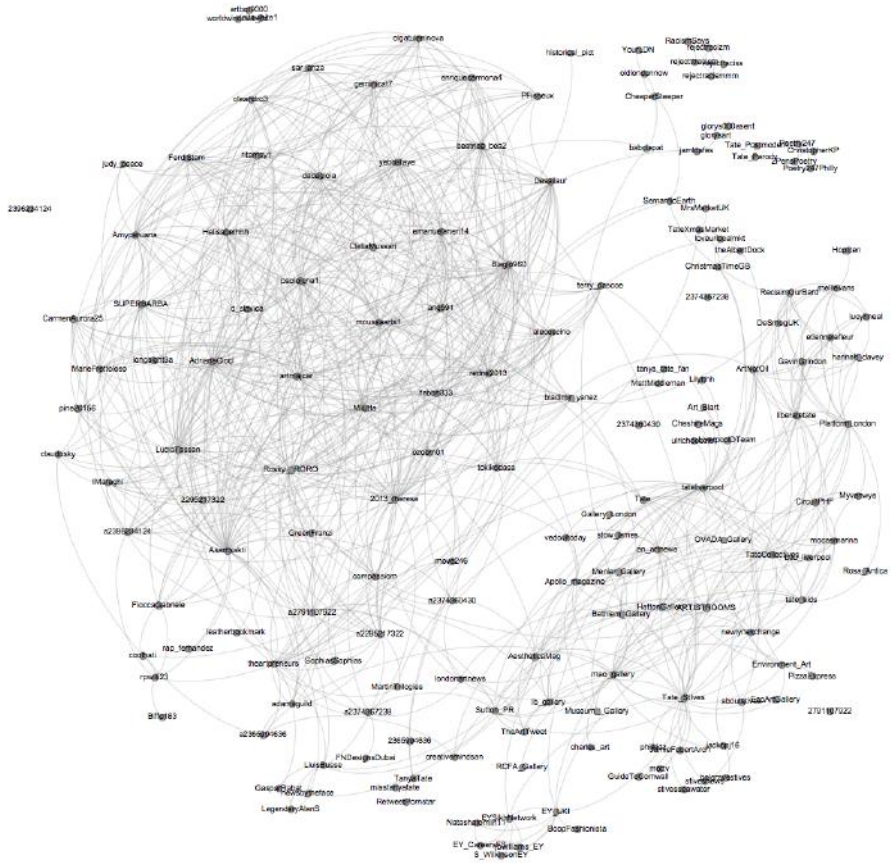

**Fig. 5-21**. Network graph of reciprocally connected Twitter users with posts including 'Tate'; note small disconnected cliques relating to unrelated networks, top right.

**Social Network Analysis Stratified by Centrality**

In our analysis of this data set we acknowledge that all users are not the same; indeed, using social network analysis we can begin to uncover the different roles within our social media network. To do this, we follow a stratified approach based on the **centrality metric** of nodes within the network: firstly identifying the most central nodes, then the wider group of most central nodes, and then finally, the least central group of nodes. We do this by filtering on the range of betweenness centrality of nodes (range=0-1938.39; shown in Fig. 5-22), starting with the first section of similar density from the point of greatest centrality (shown in the extreme right area of the x-axis). Using this technique (specifically the 'scree slope' centrality frequency distribution filter function in Gephi) we generate the following groupings used in the following analysis: Most central nodes (from maximum centrality of 1938.4 to lower bound 814.1); Most central including a wider range of nodes (again from the maximum centrality, but with a lower bound of 97.4); Least central nodes (from the minimum centrality of 0 to the maximum bound of 97.4).

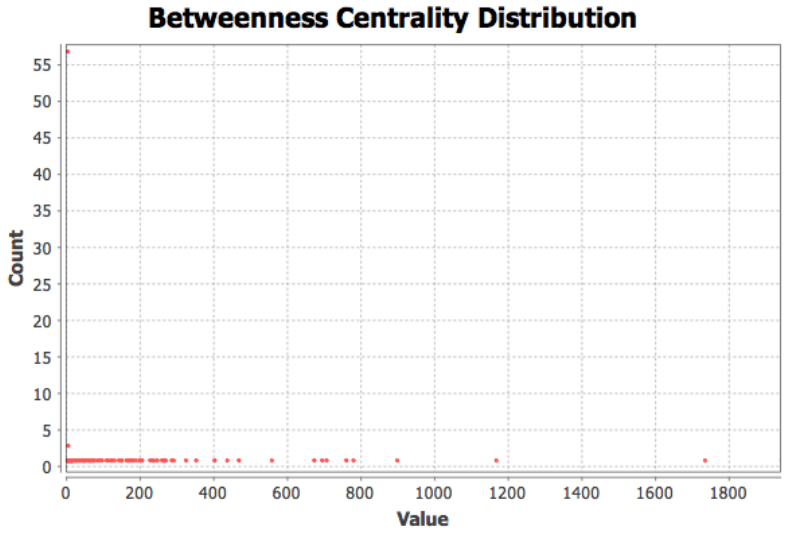


**Fig. 5-22**. Betweenness centrality distribution of reciprocally connected Twitter users with posts including 'Tate'.

For the most central nodes we identify four: 'tateliverpool', 'Tate_StIves', and 'mao_gallery' in a cluster, along with a sole node, 'Asamsakti' (betweenness centrality range filtered between 814.1 and the maximum of 1938.4). Although all nodes are in some way influential in the network, our analysis has identified two different clusters representing two types of participant within the network: the three most central nodes forming a cluster are Twitter accounts relating to galleries, two of which are part of Tate (Liverpool and St Ives), along with the Modern Art Oxford gallery (mao_gallery); in addition, there is also the single node ('Asamsakti') which belongs to an influential individual, describing himself as (among other things) an 'art lover', and who regularly tweets and retweets art images or stories that are of interest to him and his followers (with respect to the galleries, we note a division in their use of Twitter, with Tate St Ives being more tweeted about, and the Tate Liverpool and MAO using Twitter to promote news and exhibitions). Note that betweenness centrality implies that a node is significant to information transfer through a network: it does not necessarily mean that the node is itself a/the primary source of information.

By extending our analysis to the next section of the centrality distribution slope, we gain a further 9 nodes, giving a total of 13 nodes (range 97.4-1938.4; with 33 edges; Fig. 5-23). In addition to the nodes already described, we also gain 'an_artnews', 'AestheticaMag', 'Apollo_magazine', and 'EY_UKI' which group with the gallery cluster, along with '2013_therese', 'mousaartx1', 'AdrianaCioci', 'bladimir_yanez', and 'geminicat7' which cluster with the previous individual 'art

lover' node ('Asamsakti'). With this broadening out of the network, we begin to understand how the different influential clusters (and nodes) relate to each other within Twitter. Within the gallery cluster we can see how the new nodes provide an interface between the galleries themselves and 'art lovers' cluster, due to the fact that they are related to publications and news services (A-N News, Aesthetica Magazine, and Apollo Magazine) which target artists and those strongly involved in the visual arts scene (we also note that Tate St Ives, to some extent, performs a bridging role), and thus acts as an information broker. A lone node linked to the two Tate galleries is EY_UKI (formerly Ernst & Young the professional services organisation), which presumably has been a corporate sponsor. The 'art lover' cluster is apparently cohesive in terms of user type; the additional five nodes use Twitter to fulfill largely the same purpose as 'Asamsakti', namely the tweeting and retweeting of art images and other stories of interest (mainly but not exclusively art-related).

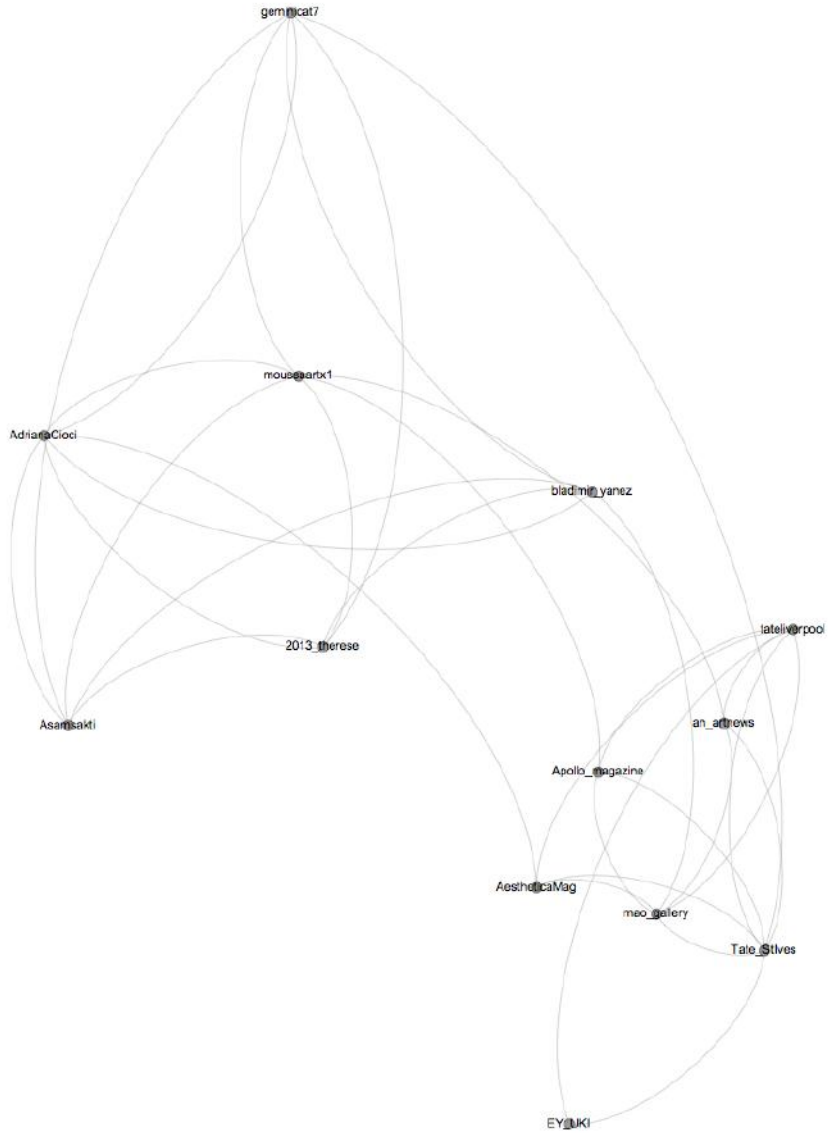


**Fig. 5-23**. Network graph of 13 most central reciprocally connected Twitter users with posts including 'Tate'.

Finally, by contrast, examining the lowest level of the scree slope of the centrality distribution gives 159 nodes and 535 edges (range 0-366.4; Fig. 5-23). Examining this in the context of the most central 13 nodes just described, it is clear that there is a large cluster of less central 'art lovers' associated with those already identified situated in the top right hand corner of the graph; similarly, more

galleries and art magazines can be seen in the middle right hand side (interestingly, here we find 'Tate', which apparently has less centrality on Twitter in comparison with either Tate St Ives or Tate Liverpool). Around the right hand side of the graph periphery, we find smaller clusters relating to the galleries, for example in terms of tourism based on location (e.g. 'GuideToCornwall'), events at Tate ('tate_kids'), EY ('EY_Careers'), or campaigning or critiquing in relation to the Tate galleries (e.g., 'liberatetate').

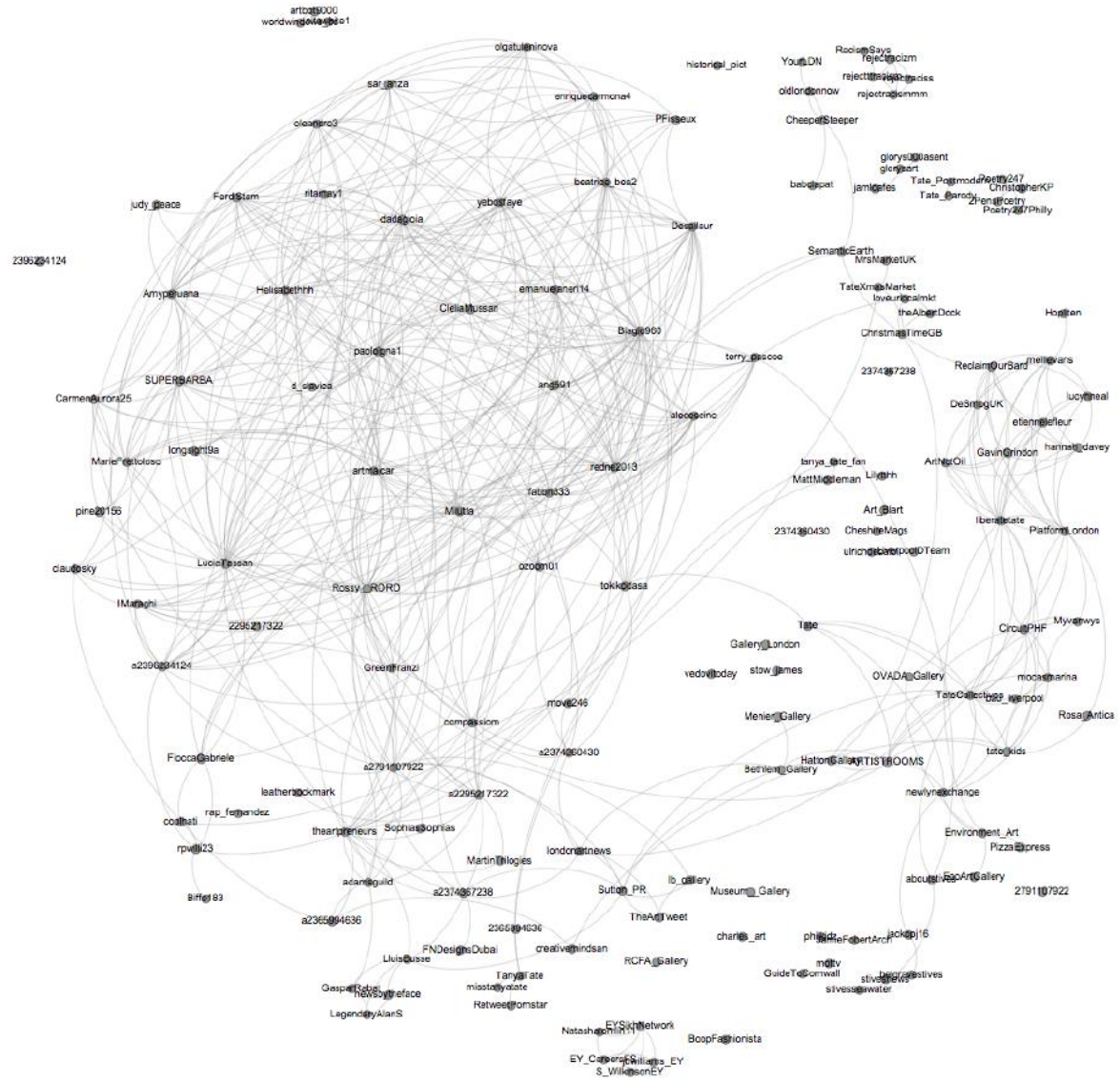


**Fig. 5-24**. Network graph of least central reciprocally connected Twitter users with posts including 'Tate'.

Referring back to the graph showing the totality of the Tate Twitter network, we can summarise it in terms of clusters of users and their functions: the most central nodes belong to the regional Tate galleries along with the Modern Art Oxford gallery (centred in the lower right area in Fig. 5-24), along with a single influential 'art lover' (upper right); when nodes with lower centrality are included, we note the greater size of the 'art lover' cluster, relative to the gallery cluster. Framing these clusters in terms of users, we can characterise the 'art lovers' as largely relating to the general public who are interested in learning about and sharing art (perhaps using Twitter as a form of scrap book); the gallery cluster relates to art institutions publicising their exhibitions, art objects, news and events. In addition to these two main clusters, we also found that art magazines and news services tended to cluster with the galleries, and while demonstrating lower centrality, they performed the important

role of providing an interface between the public 'art lover' users and the galleries themselves, digesting, filtering, aggregating and promoting relevant features for their audience (both other galleries and the public). Finally, we note a range of small clusters with lower levels of centrality which tag around the periphery of the graph, in some cases promoting specific Tate events, others protesting about commercial sponsorship of Tate, with others relating to tourism or corporate sponsorship.

Translating this analysis into personae that could represent the social media user community, we identify four main groups:

- Art institution marketing team (internal)
- Other art institutions
- Art news services
- Interested general public

In addition, we also identify smaller groups of users, which are:

- Tourism agencies related to the institutions (external; presumably internal requirements will be met by the art institution marketing team, already mentioned above)
- Corporate sponsors
- Campaigning groups

#### *SUMMARY AND CONCLUSION*

We have utilised two approaches to understand the user community around art institutions. Focusing in particular on the Tate, we have examined the semantic content of catalogue titles of the Tate in relation to a wider selection of art museums, and have explored the social network present in the Twitter user community around Tate. The former approach examined the art catalogue as a lens utilised by archivists and curators which represented their perceptions of their user community.

From this analysis, we were able to identify emergent semantic themes most relevant to the user community around the Tate. These were: humans, nature, the outdoors, and particular places. In our latter analysis, we concentrated on a particular subsection of social media, specifically reciprocal links within Twitter-users who post content about Tate. From this analysis, we identified four main groups of users: art institution marketing team, other art institutions, art news services, and interested members of the general public. In addition, smaller clusters at the fringes included tourism agencies, corporate sponsors, and campaigning groups.

In future work we anticipate utilising this information to identify relevant user communities or user community change, for example by analysing changes to the social network relating to a cultural institution (we address this in T5.3.3). The key datasets from this work may be found at http://seis.bris.ac.uk/~cselt/datasets.html.

### *5.3.4. Regularized Topic Models*

This section contains guest analytical work fitting D4.4, but not included in the methodological spectrum of PERICLES. We are grateful to Artem Popov (Lomonosov Moscow State University), Anna Potapenko (National Research University Higher School of Economics), and Konstantin Vorontsov (Moscow Institute of Physics and Technology, Dorodnicyn Computing Centre of RAS) for their contribution and permission to include their results. The aim is to identify topic shifts by a probabilistic (i.e. neither vector- nor graph-based) method.

#### *INTRODUCTION*

In the modern world there is a lot of information to store and organize. This information can be effectively researched by statistical methods. The goal of such research is data insights, which can be

useful for applied fields. In this work we explore the collection of 46381 images from Tate gallery (investigated in the previous subsection as well) based on their catalogue metadata by BigARTM, a topic analysis software tool[64]. As regards metadata, and as already mentioned previously, each art object is manually annotated by labels of three levels – from more general concepts to more specific objects in the paintings. Every art object also has a time-stamp: 33625 pictures come from 1800s (1796–1845) and 12756 pictures from 2000s (1960–2009). Each period is further divided into 10 five-year frames called epochs previously in this deliverable.

The aim of the research is to capture how art patterns change over time through statistical analysis of annotations. We are interested in obtaining a set of common topics by clustering the labels and then detecting topic shifts, i.e. capturing how those clusters evolve from one period to another. Labels of the first and second levels are too few for a comprehensive statistical analysis, therefore we use the third level labels (7070 for the 1800s series and 6680 for the 2000s series). Further labels are referred to as words and annotations as texts.

### *Probabilistic Topic Modelling*

**Topic modelling** is a widely used approach for soft bi-clustering of words and texts [Blei, 2012]. Given corpus of texts and a number of topics, it learns a multinomial distribution over words for each topic, and then describes each document with a multinomial distribution over topics. Such representation reveals a hidden thematic structure of the collection and promotes the usage of topic models in information retrieval, classification, categorization, summarization and segmentation of texts.

#### Incorporating Time-stamps

The goal of our research implies that we are looking not only for topic distributions over words, but we also need time-stamps in the model to trace how those distributions evolve over time. Topics Over Time [Wang & McCallum, 2006] is one of the first models, which considers time-stamps as pseudo-words and learns a distribution over time-stamps for each topic. Dynamic Topic Model [Blei & Lafferty, 2006] is another popular approach that learns separate topic models for each period but makes sure to keep them similar by using priors.

#### Processing Short Texts

Another important requirement concerns the shortness of texts (each picture is annotated by a few labels), which is usually an issue in topic modelling because word-document matrix becomes too sparse. Biterm Topic Model [Cheng et al., 2014] explicitly models pairs of words (biterms) that co-occur in a document. Other successful approaches [Yan et al., 2013; Zuo & Xu, 2014] also utilize word-word co-occurrence statistics to build a topic model.

### *Review of Existing Models*

Although a lot of topic models have been developed for different tasks, combining several requirements in one model remains an open problem. The most popular topic models are Probabilistic Latent Semantic Analysis [Hofmann, 1999] and its Bayesian extension called Latent Dirichlet Allocation [Blei et al., 2003]. Incorporating a new prior to build a topic model that meets the certain requirements might lead to a complicated and sometimes unfeasible inference. Therefore, we develop a non-Bayesian approach of additive regularization that removes a lot of limitations and simplifies theory without loss of generality [Vorontsov & Potapenko, 2015].

[64] The development of BigARTM (http://bigartm.org/) was led by Oleksandr Frei (Schlumberger Information Solutions). An introduction to its use can be found in [Vorontsov et al., 2015a].

### ADDITIVE REGULARIZATION OF TOPIC MODELS

Let $D$ denote a finite set (collection) of documents (texts) and let $W$ denote a finite set (vocabulary) of all terms from these documents. Following the "bag of words" hypothesis, we represent each document $d$ from $D$ as a subset of terms from the vocabulary $W$ with the respective frequencies denoted by $n_{dw}$.

A probabilistic topic model represents the probabilities $p(w|d)$ of terms occurring in documents as mixtures of term distributions in topics $\phi_{wt} = p(w|t)$ and topic distributions in documents $\theta_{td} = p(t|d)$:

$$p(w|d) = \sum_t p(w|t)p(t|d) = \sum_t \phi_{wt}\theta_{td}$$

Parameters of a topic model are represented as matrices $\Phi = (\phi_{wt})$ and $\Theta = (\theta_{td})$ with non-negative and normalized columns $\phi_t$ and $\theta_d$ representing multinomial word-topic and topic-document distributions respectively.

In Additive Regularization of Topic Models (ARTM) [Vorontsov & Potapenko, 2015] a topic model is learned by maximization of a linear combination of the log-likelihood $L(\Phi, \Theta)$ and $r$ regularizers $R_i(\Phi, \Theta)$, with regularization coefficients $i = 1, 2 \dots r$ given non-negativity and normalization constraints:

$$\underbrace{\sum_{d,w} n_{dw} \log \sum_{\mathrm{t}} \phi_{wt}\, \theta_{td}}_{L(\Phi,\Theta)} + \underbrace{\sum_i \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta} \quad ,$$

Learning a topic model is an ill-posed problem of approximate stochastic matrix factorization, which has an infinite set of solutions. Regularization penalty terms help to choose a better solution and can be added in any combinations. In this research we use the following:

- Decorrelating regularizer minimizes the sum of pair-wise covariances between $\phi_t$ vectors for topics (make topics more diverse, stimulates sparsity and tends to group stop-words and common words into separate topics).
- Sparsing regularizer maximizes Kullback-Leibler divergence of a given and uniform distribution (encourages topics to be concentrated on a relatively small subset of words and documents and assignes each document to a few topics).
- Topic selection regularizer starts with excessive number of topics and removes less significant and linearly dependent topics (helps in optimizing a number of topics).

### PROPOSED MODELS FOR TATE DATA

Topic models were built for 1800s and 2000s data individually in a similar fashion. Raw input data were processed to form a word-word matrix, where each cell corresponds to a number of documents, which contain both words. This co-occurrence matrix is further factorized to obtain $\Phi$ and $\Theta$ matrices during learning process. In terms of previous sections, documents are now substituted by aggregated contexts of words, i.e. we consider $|W|$ pseudo-documents, each formed by concatenation of all initial documents where a certain word $w$ occurs. This addresses the problem of short texts.

To address the concept of evolving topics, we firstly learn a general model on the whole data and then use it as an initialization to fine-tune on each period separately. To obtain a better interpretability of topics, these design solutions are combined with manually adjusted regularization strategy that consists of three stages:

1. Intense decorrelating to make topics as diverse as possible.
2. Alternating decorrelating steps and topic selection to discard insignificant topics.

3. Incorporating Φ-sparsity regularizer to make topics fine-grained.

Detailed information about regularization coefficients is presented in Table 5-8. To make sure that a topic model covers most of significant topics and choose a number of_topics we followed the following procedure. Several topic models with different regularizers and_number of topics were built. Then a set of interpreted topics, which have ever encountered in a model, was manually created. A final model was required to contain each topic from this set.

**Table 5-8**. Regularization strategy: adjusted coefficients.

| Period | Stage | Num.iter. | Decorrelation | Topic select. | Phi sparsity |
|---|---|---|---|---|---|
| 1800s | 1 | 10 | 7590 | 0 | 0 |
| | 2 | 25 | 7590 / 0 | 0 / 0.273 | 0 |
| | 3 | 10 | 7590 | 0 | -0.001 |
| 2000s | 1 | 25 | 1565 | 0 | 0 |
| | 2 | 25 | 1565 / 0 | 0 / 0.3484 | 0 |
| | 3 | 10 | 1565 | 0 | -0.005 |

## *Experiments and Results*

The approach of additive regularization is implemented in open-source library of topic modelling BigARTM.org [Vorontsov et al., 2015a]. It was used for building all models during this research.

When building topic models, we focused on both quantitative and qualitative analysis of the results. Fig. 5-25 presents dependence of several quality measures on iterations of a learning algorithm. Perplexity Score (the lower the better) is based on likelihood of the model and is widely used as a primary quality measure in topic modelling. Sparsity shows a ratio of zero elements in obtained distributions and captures the intuition of fine-grained and specific topics.

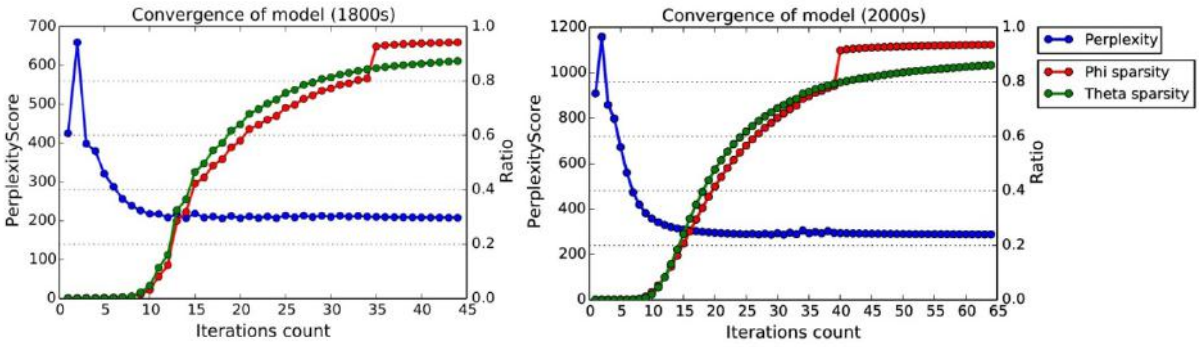


**Fig. 5-25**. Convergence charts.

To investigate and interpret the topics and topic shifts we built several types of results representations[65]:

- Top-words for topics (built on full data) and for topics in periods (fine-tuned for each period independently)
- Labels probabilities for each period for each topic (how word structure evolves in time)
- Topic popularity on a year (how topics themselves evolve in time)

To organize the obtained results and provide some analysis we divide all interpretable topics into several groups based on how the representations of topics in periods correspond to each other and to the global representation.

[65] Can also be found at https://www.dropbox.com/sh/pjwdtqx7oz8gku2/AADdnXp4Kd_k8IrZRBddid-Ja?dl=0

**Constant Topic, Constant Vocabulary**

These topics stay constant during all periods both in terms of word structure and their overall popularity. Usually these topics are devoted to nature, sea, architecture or interior. Some examples are presented below. The left chart of Fig. 5-26 shows how popularity of a constant topic depends on year and the top chart of Fig. 5-27 shows how popularity of labels changes in periods (the darker the more).

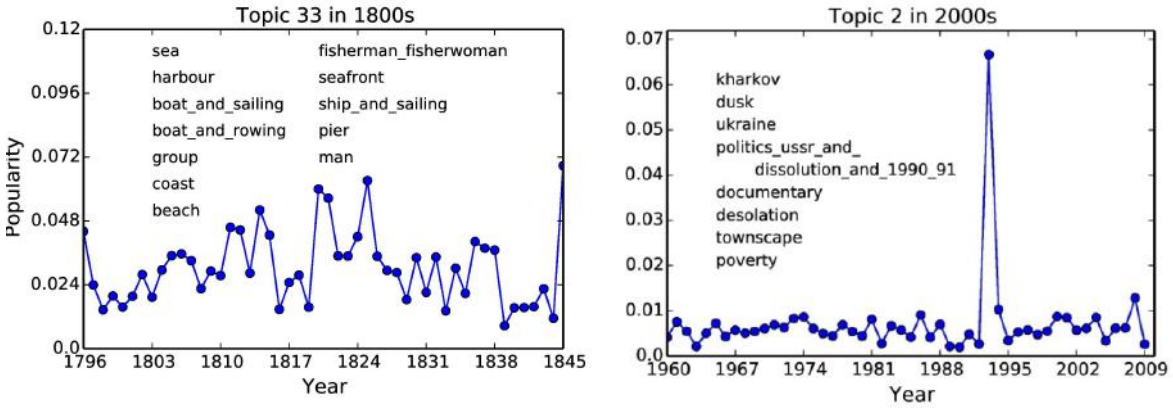


**Fig. 5-26**. Topics in time: constant topic, constant vocabulary (left) and event-related topic (right).

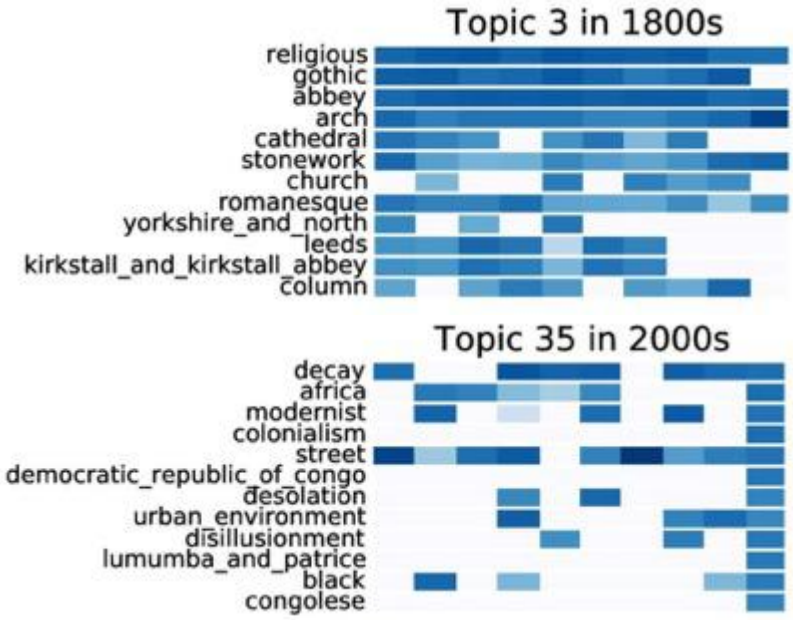


**Fig. 5-27**. Words in time: constant topic, constant vocabulary (top) and event-related topic (bottom).

**Event-related Topics**

These topics are concentrated and well interpretable only in one period and correspond to some events either in world or art that happen in that time. These topics are often related to wars, politics or racism. For example, the right chart of Fig. 5-26 presents a topic about USSR dissolution and has the greatest popularity in the corresponding years.

**Constant Topic, Dynamic Vocabulary**

Probably, it's the most unusual and interesting type of topics. They stay interpretable during all the periods and capture one general concept, but describe this concept in different words from period to

period (for example, words related to sequence of historical events). Table 5-9 shows top-words of such a topic in several periods. In general, the topic is about war and aggression, but in different periods it captures different related events – war in Vietnam, World War II or Bosnian War. So we can clearly observe how key words for one and the same topic evolve over time.

**Table 5-9**. Shifts in key words for a topic devoted to war in 2000s.

| | |
|---|---|
| Period 1 | skull, antihero, war, blood, carrying, destruction, fire, menace, bandage, flagpole, revolution, soldier, pain, nazism_swastika, screaming |
| Period 2 | blood, lying_down, war, horror, protests_and_unrest_anti_vietnam_war_demonstration_and_may_1970, vietnam_war_and_1964_75, militarism, wheelchair, disability |
| Period 3 | murder, horror, corpse, mutilated, torture, death, soldier, war, goya_and_francisco_de_and_etching_and_disasters_of_war, bleeding, artillery, bosnian_conflict_and_1990s, bosnia_and_hercegovina |

### Long-lived Topics

These topics are neither constant, not event-related and stay popular and interpretable in a number of periods (3 or more). Popularity dependence on time is usually multimodal. The standard subjects would be nature, science, concrete personalities, concrete places. See Table 5-10 for top-words of some topics of this type and also all the others to get more examples.

**Table 5-10**. Most probable words (labels) for several topics of different types.

| 1800-6 (long-lived) | 2000-50 (long-lived) | 1800-0 (constant) | 2000-17 (constant) |
|---|---|---|---|
| canal | film_disney_and_walt | mountain | interior |
| venice | mouse | lake | window |
| waterfront | mickey_mouse | alps | curtain |
| boat_and_gondola | cultural_icon | switzerland | chair |
| venice_and_grand_canal | donald_duck | wooded | table |
| venice_and_doge_s_palace | duck | valley | door |
| dome | cartoon_comic_strip | rocky | shadow |

| 2000-40 (long-lived) | 2000-7 (event-related) | 2000-32 (event-related) | 2000-87 (event-related) |
|---|---|---|---|
| robot | politics_argentina_1976_1983 | political_protest | suffering |
| astronaut | human_rights | commerce | persecution |
| science | totalitarianism | defacement | holocaust |
| spacecraft | argentina | money | victim |
| helmet | videla_and_jorge_rafael | corruption | torture |
| pilot | politician_and_president | politics_brazil | male |
| science_fiction | massera_and_emilio_eduardo | political_prisoner | jewish |

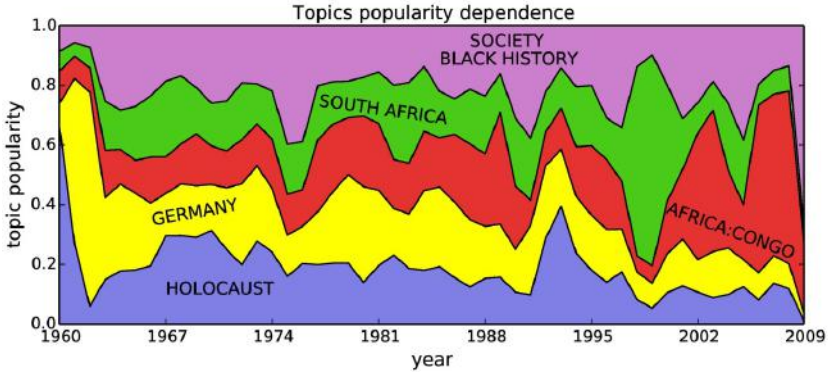


**Fig. 5-28**. Topics popularity dependence.

Until now we have analysed several topics independently, but it is also interesting to discover, how they relate to each other. Particularly, we would be interested in finding interpretable correlation of topics behaviour in time. One example of such correlation is presented in Fig. 5-28. Here a topic about German and a topic about Jewish have large popularity in the same period of time (1960 – 1963) when probably related pieces of art were created.

### *SUMMARY*

In this research we have analysed Tate data with additively regularised topic models and revealed some interpretable topics and their dependencies. We have traced how topics evolve over time in terms of their key words and popularity. The analysis show that pieces of art that date back in 1800s are mostly devoted to nature and are constant, while 2000s mostly present politics or interior topics and are often event-related in rapidly changing over time.

## 5.4. Chapter Summary

Chapter 5 focused on semantic change, an important research problem within PERICLES, which monitors and measures changes in the meaning of concepts within knowledge representation models. As discussed, this phenomenon involves **technological obsolescence** along with **changes in language and society** and, thus, bears drastic consequences in LTDP. Due to its multifaceted profile, semantic change has been the subject of various lines of research attempting to tackle the associated challenges. In this chapter, we presented the novel approaches we adopted for studying and monitoring semantic change: (a) **a vector field approach** to evolving semantics, (b) **an ontology evolution approach** for deploying concept drift principles to semantic change, and, (c) a methodology for **studying semantic change in communities**. All the presented approaches are accompanied by respective open-source software tools that will be integrated in the PERICLES testbeds. Finally, a fourth "guest" approach was also presented in this chapter driven by partners outside the PERICLES consortium using a probabilistic method to identify topic shifts in an evolving collection. These four approaches complement one another and provide digital curators with a powerful toolkit for collection and collection use diagnostics.

# 6. Conclusions and Next Steps

## 6.1. Conclusions

The deliverable reported on the work conducted in T4.4, focusing on the modelling of content and context and on contextualised content semantics, describing our two diverse but complementary approaches: ontology-based and statistical. More specifically, the following outputs per topic were presented:

- **Ontology-based Representation of Content and Context**: The deliverable presented our proposed ontology-based models for semantically representing content items and their context, including use-context (i.e. context of use of a content item). The adopted formalisms for the Art & Media and the Space Science domain are OWL and Topic Maps, respectively. As elaborated in the respective chapter, although the two domains are vastly different, the developed ontologies share several common characteristics, which are inherited by the adopted LRM on which the ontologies are based on. The document also discussed our proposed contextualised semantics methodologies for taking advantage of context representation. In this direction, we have deployed an additional inference layer on-top of the developed OWL models, which is based on the SPARQL Inferencing Notation (SPIN) and which efficiently handles context-related inconsistencies. Several sample implementations and indicative examples from the domain ontologies were also presented.
- **Statistical Context Modelling and Contextualised Content Semantics**: After having introduced contextualised content semantics in the intellectual framework of ontology construction and development, the deliverable also presented our second approach for modelling context, treating context-dependent, evolving semantic content as a vector field by a combination of ideas from linguistics, statistics and classical mechanics, based at its core on multivariate statistics for scalability, tool and methodology testing. We evaluated the feasibility of our considerations on a major text dataset and image metadata from the online catalogue of Tate Gallery. For these analyses, we used Somoclu, another core PERICLES outcome that is a high-performance qualitative machine learning algorithm suitable for exploratory data analysis.
- **Semantic Change and Evolving Semantics**: Based to a great extent on the proposed representations for content and context, the deliverable studied semantic change along with the overall phenomenon of evolving semantics and presented our three lines of investigation in this area:
  - A field approach to evolving semantics, dealing with textual content and indexing terminology change, based on the theory of semantic fields, blended with multivariate statistics and the concept of fields in classical mechanics to enable machine learning. A proof-of-concept tool for detecting and measuring semantic drifts has been developed with micro- and macroscopic analytical abilities, i.e. zooming in and out, and taking snapshots of content distributions at regular intervals to record changes. The tool can scan scalable sets of objects and/or features to inspect trends on the deepest level given by e.g. an index term hierarchy.
  - A study of semantic change under an ontology evolution perspective, investigating changes occurring in ontology models. The adopted methodology, stemming from existing work in concept drift, measures semantic drift considering (a) the different aspects of change, and, (b) whether concept identity is known or not. The different types of change, reflecting a concept's meaning, include the label, intension and extension. Additionally, the correspondence of a concept across versions of an ontology can be either known or unknown, resulting respectively in the identity-based and morphing-based approaches for

measuring change. We also developed an open, reusable software solution that adopts, extends and implements these methods, presented in the deliverable.

- A study of community change in social media, utilising two approaches to understand the user community around art institutions. We examined the semantic content of catalogue titles of the Tate in relation to a wider selection of art museums, and explored the social network present in the Twitter user community around Tate. From this analysis, we were able to identify emergent semantic themes most relevant to the user community around the Tate, like e.g. humans, nature, the outdoors, and particular places. In our analysis, we also concentrated on a particular subsection of social media, specifically reciprocal links within Twitter-users who post content about Tate. From this analysis, we identified the main groups of users.

- Besides the above three approaches, the deliverable also presented a fourth "guest" line of analytical work on **topic shifts** by affiliated partners outside the project's consortium. In this research Tate data was analysed with additively regularised topic models and revealed some interpretable topics and their dependencies. The way topics evolve over time in terms of their keywords and popularity was traced and the analysis showed that pieces of art that date back in 1800s are mostly devoted to nature and are constant, while 2000s mostly present politics or interior topics and are often event-related in rapidly changing over time.
- All **developed models and software tools** for the investigations within D4.4 are publicly available along with the respective results and datasets.

## 6.2. Next Steps

This subsection discusses the main directions for follow-up on the activities reported in this deliverable, either for the upcoming PERICLES tasks (e.g. T4.5 and tasks within WPs 3, 5, 6) or even beyond the lifetime of the project, thus, securing the sustainability of our research findings.

**Links to upcoming PERICLES tasks**

Regarding the ontology-based models for semantically representing content and context (see Chapter 3), there are still a few refinements left to integrate, according to ongoing requirements emerging from work with the WP6 testbeds. Besides the already proposed ODP from the DVA domain (see Section 3.1.3), a couple of additional ODPs are planned to be submitted to relevant venues, describing core aspects of the other two Art & Media subdomains, SBA and BDA.

Additionally, work on contextualised semantics is still ongoing and will feed into the upcoming T4.5. Further, the presented inference layer (see Section 3.3) will be supplemented with more powerful reasoning mechanisms forming a semantic interpretation framework for the high-level integration and semantic fusion of content and context knowledge. Approaches for handling inconsistent/missing knowledge will also be considered.

Furthermore, the evolving semantics investigations presented in Chapter 5 are highly adaptable and will result in generic tools for detecting and measuring semantic change, applicable in any scenario within and outside PERICLES. However, certain aspects will feed into WP3 and WP5 tasks. More specifically:

- Vector field output by Somoclu in general and semantic drift metrics in particular feed to T3.5.4 and T5.3.3, with the open source tool submitted to WP6 for deployment.
- The ontology evolution metrics introduced in Section 5.3.2 will feed into the DEM ontology within the context of T3.5.
- The methodology and tools for detecting user community change (Section 5.3.3) will be deployed in analysing changes to the social networks relating to cultural heritage and memory organisations (addressed in T5.3.3).

With physics as a metaphor for evolving semantics, in T4.5.1 we continue tests from classical mechanics and quantum theory to better understand the features, classes, behaviour and use context of digital objects important for DP.

Additionally, based on using RDF statements as terms for the indexing of DOs, the LRM and the field model can be gradually merged and a next curation tool designed, combining ontology-based feature definition with vector field semantics. Such a new experiment in the arts domain is in progress and its output will be relevant both to T5.3.3 and, given its high relevance to the Semantic Web, to research beyond PERICLES.

**Links beyond PERICLES**

The research activities within WP4 are highly novel and cutting-edge and, thus, their applicability is not exhausted only in the confines of the PERICLES project. Instead, several of the lines of research can be extended outside the scope of the project. For instance, regarding the methods developed for calculating drift measures based on ontology evolution (see Section 5.3.2), their deployment outside the scope of PERICLES involves implementing the “identity-based” method as well. Having both the “morphing-based” and the “identity-based” methods can lead to implementing a cross-platform full-fledged software tool that will allow the end-user to interlink concepts across ontology versions using graphical means. Further, we have plans to combine Somoclu and BigARTM developed by our Russian research partners and to study the parallel evolution of tension vs. content structure in a vector field.

# 7. References

[Abel et al., 2012] Abel, F., Hauff, C., Houben, G. J., Stronkman, R., & Tao, K. (2012). Semantics + Filtering + Search = Twitcident. Exploring Information in Social Web Streams. In Proceedings of the 23rd ACM conference on Hypertext and Social Media, ACM, pp. 285-294.

[Achilleos et al., 2010] Achilleos, A., Yang, K., Georgalas, N. (2010). Context modelling and a context-aware framework for pervasive service creation: A model-driven approach. Pervasive Mob. Comput. 6, pp. 281–296.

[Allen, 1983] Allen, J.F. 1983. Maintaining knowledge about temporal intervals. Communications of the ACM. 26 November 1983. ACM Press. pp. 832–843.

[Allinson, 2006] Allinson, J. (2006). OAIS as a reference model for repositories: an evaluation. UKOLN, University of Bath.

[Arndt et al., 2007] Arndt, R., Troncy, R., Staab, S., Hardman, L., and Vacura, M. (2007). COMM: Designing a Well-founded Multimedia Ontology for the Web. Proc. 6th Int. Semantic Web and 2nd Asian Semantic Web Conf. (pp. 30-43). Busan, Korea: Springer-Verlag.

[Baader et al., 2003] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D. and Patel-Schneider, P.F. (2003). The Description Logic Handbook. Cambridge University Press.

[Baldauf et al., 2007] Baldauf, M., Dustdar, S., and Rosenberg, F. (2007). A survey on context-aware systems. Int. J. Ad Hoc Ubiquitous Comput. 2, pp. 263–277.

[Baroni et al., 2007] Baroni, M., Lenci, A., and Sahlgren, M. (Eds.) (2007). Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models Beyond Words and Documents. Computer Science Research Report No. 116, Roskilde University: Roskilde.

[Baroni & Lenci, 2010] Baroni, M., Lenci, A. (2010). Distributional memory: A general framework for corpus based semantics. Computational Linguistics 36(4), pp. 673-721.

[Bennett & Galton, 2004] Bennett, B. and Galton, A.P. (2004). Unifying events and time. Artificial Intelligence 153, pp. 13-48.

[Bettini et al., 2010] Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., and Riboni, D. (2010). A survey of context modelling and reasoning techniques. Pervasive Mob. Comput. 6, pp. 161–180.

[Blacoe et al., 2013] Blacoe, W., Kashefi, E., Lapata, M. (2013). A Quantum-Theoretic Approach to Distributional Semantics. In Proceedings of NAACL-HLT, pp. 847–857.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

[Blei & Lafferty, 2006] Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120). ACM.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of machine Learning research, 3, 993-1022.

[Bloehdorn et al., 2004] Bloehdorn, S., Simou, N., Tzouvaras, V., Petridis, K., Handschuh, S., Avrithis, Y., Kompatsiaris, I., Staab, S., and Strintzis, M. (2004). Knowledge Representation for Semantic Multimedia Content analysis and Reasoning. Proc. of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT), London, U.K., November 25-26.

[Charles & Isaac, 2015] Charles, V., and Isaac, V. (2015). Enhancing the Europeana Data Model (EDM). Available online at: http://pro.europeana.eu/publication/enhancing-the-europeana-data-model-edm.

[Chen et al., 2005] Chen, H., Finin, T., and Joshi, A. (2005). The SOUPA ontology for pervasive computing, in: Ontologies for Agents: Theory and Experiences. Springer, pp. 233–258.

[Cheng et al., 2014] Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. Knowledge and Data Engineering, IEEE Transactions on, 26(12), 2928-2941.

[Chi, 2015] Chi, Y. (2015). A Complete Assessment of Tagging Quality: A Consolidated Methodology. JASIST 66(4), pp. 798-817.

[Chou et al., 2005] Chou, S.-C., Hsieh, W.-T., Gandon, F.L., and Sadeh, N.M. (2005). Semantic web technologies for context-aware museum tour guide applications, in: Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on. IEEE, pp. 709–714.

[Coecke et al., 2010] Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical Foundations for a Compositional Distributional Model of Meaning, at http://arxiv.org/abs/1003.4394

[Community Systems Group, 2007] Community Systems Group (2007). Community systems research at Yahoo!. SIGMOD Rec. 36, 3, pp. 47-54.

[Couch, 1992] Couch, C. J. (1992). Toward a formal theory of social processes. Symbolic Interaction 15(2), pp. 117-134.

[Dallas, 2007] Dallas, C. (2007). An agency-oriented approach to digital curation theory and practice. In, J. Trant and D. Bearman (Eds), Proceedings of the International Cultural Heritage Informatics Meeting (ICHIM07). Toronto: Archives and Museum Informatics, at http://www.archimuse.com/ichim07/papers/dallas/dallas.html

[Darányi & Wittek, 2013a] Darányi, S. and Wittek, P. (2013). Connecting the Dots: Mass, Energy, Meaning, and Particle-Wave Duality. In Proceedings of QI-12. Springer.

[Darányi & Wittek, 2013b] Darányi, S. and Wittek, P. (2013). Demonstrating conceptual dynamics in an evolving text collection. Journal of the Association for Information Science and Technology 64(12), pp. 2564–2572.

[Dasiopoulou et al., 2007] Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., and Strintzis, M. (2007). Capturing MPEG-7 Semantics. Proc. 2nd International Conference on Metadata and Semantics (MTSR), Corfu, Greece.

[Dasiopoulou et al., 2010] Dasiopoulou, S., Tzouvaras, V., Kompatsiaris, I., and Strintzis, M. (2010). Enquiring MPEG-7 based Multimedia Ontologies. Multimedia Tools Appl. 46, 2-3 (January 2010), pp. 331-370.

[Dekker et al., 2015] Dekker, A., Falcão, P. and Laurenson, P. (2015). An exploration of significance and dependency in the conservation of software based artworks, AIC's 43rd Annual Meeting - Electronic Media Session.

[Denbigh, 1982] Denbigh, K.G. (1982). Three concepts of time. Springer: Berlin.

[Dey et al., 2001] Dey, A. K., Abowd, G. D., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Human-computer interaction, 16(2), 97-166.

[Doerr, M., 2005] Doerr, M. (2005). The CIDOC CRM, an Ontological Approach to Schema Heterogeneity. Dagstuhl Seminar Proceedings, No. 4391 - Semantic Interoperability and Integration.

[Doerr & Theodoridou, 2011] Doerr, M. and Theodoridou, M. (2011), CRMdig: A generic digital provenance model for scientific observation. TaPP' 11, 3rd USENIX Workshop on the Theory and Practice of Provenance, Heraklion, Crete, Greece

[Fanizzi et al., 2008] Fanizzi, N., d'Amato, C. and Esposito, F. (2008). Conceptual clustering and its application to concept drift and novelty detection. Springer.

[Fellbaum, 1998] Fellbaum, Ch. (Ed.) (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge MA.

[Firth, 1957] Firth, J.R. (1957). Papers in Linguistics 1934–1951 (1957) London: Oxford University Press.

[Frommholz et al., 2010] Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., and van Rijsbergen, K. (2010). Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In Proceedings of the third symposium on Information interaction in context (IIiX '10), pp. 115-124. ACM: New York. DOI:10.1145/1840784.1840802.

[Gangemi, 2005] Gangemi, A. (2005). Ontology design patterns for semantic web content. In The Semantic Web–ISWC 2005 (pp. 262-276). Springer Berlin Heidelberg.

[Gangemi et al., 2002] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. and Schneider, L. (2002). Sweetening Ontologies with DOLCE. In A. Gómez-Pérez and V. Benjamins (Ed.), Proc. 13th Int. Conf. on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web (pp. 166-181). London, UK: Springer-Verlag.

[Garcia & Celma, 2005] Garcia, R. and Celma, O. (2005). Semantic Integration and Retrieval of Multimedia Metadata. Proc. International Semantic Web Conference (ISWC), Galway, Ireland.

[Goodall et al., 2003] Goodall, S., Grimwood, P., Kim, S., Lewis, P., Martinez, K., Stevenson, A., Addis, M., and Boniface, M. (2003). Sculpteur: Towards a New Paradigm for Multimedia Museum Information Handling. Int. Semantic Web Conf. (ISWC), Sanibel Island, FL, USA, pp. 582-596.

[Grefenstette et al., 2013] Grefenstette, E., Dinu, G., Zhang, Y.Z., Sadrzadeh, M., Baroni, M. (2013). Multi-step regression learning for compositional distributional semantics, available online: http://arxiv.org/abs/1301.6939

[Gu et al., 2004] Gu, T., Wang, X.H., Pung, H.K., and Zhang, D.Q. (2004). An ontology-based context model in intelligent environments, in: Proceedings of Communication Networks and Distributed Systems Modeling and Simulation Conference. pp. 270–275.

[Guha & McCarthy, 2003] Guha, R., and McCarthy, J. (2003). Varieties of contexts. In P. Blackburn et al. (Eds.): CONTEXT 2003, LNAI2680, pp. 164-177. Springer: Berlin.

[Gulla et al., 2010] Gulla, J. A., Solskinnsbakk, G., Myrseth, P., Haderlein, V., and Cerrato, O. (2010). Semantic Drift in Ontologies. In WEBIST (2) (pp. 13–20).

[Hájek, 2012] Hájek, A. (2012). Interpretations of Probability. In E.N. Zalta, (Ed.), The Stanford Encyclopedia of Philosophy, at http://plato.stanford.edu/archives/win2012/entries/probability-interpret/

[Harit et al., 2006] Harit, G., Chaudhury, S., and Ghosh, H. (2006). Using Multimedia Ontology for Generating Conceptual Annotations and Hyperlinks in Video Collections. Proc. 2006 IEEE/WIC/ACM Int. Conf. on Web Intelligence (pp. 211-217). IEEE Computer Society.

[Harris, 1968] Harris, Z. (1968). Mathematical structures of language. Interscience Publishers.

[Hollink & Worring, 2005] Hollink, L., and Worring, M. (2005). Building a Visual Ontology for Video Retrieval. In Proc. 13th ACM Int. Conf. on Multimedia, Singapore, pp. 479-482, Nov. 6-11.

[Hofmann, 1999] Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). ACM.

[Horridge et al., 2006] Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., and Wang, H. H. (2006). The Manchester OWL Syntax. OWL Experiences and Directions Workshop (OWLED) 2006, Vol. 216.

[Hu et al., 2003] Hu, B., Dasmahapatra, S., Lewis, P. H., and Shadbolt, N. (2003). Ontology-based Medical Image Annotation with Description Logics. Int. Conf. on Tools with Artificial Intelligence (ICTAI), Sacramento, California, Nov. 3-5.

[Hudelot & Thonnat, 2003] Hudelot, C., and Thonnat, M. (2003). A Cognitive Vision Platform for Automatic Recognition of Natural Complex Objects. In ICTAI, pp. 398-405.

[Hunter, 2001] Hunter, J. (2001). Adding Multimedia to the Semantic Web: Building an MPEG-7 Ontology. Proc 1st Semantic Web Working Symposium (SWWS'01), Stanford University, California, USA.

[ICA, 2000] ICA - International Council on Archives (2000), ISAD(G): General International Standard Archival Description, Second Edition, available online: http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html

[Iivari & Linger, 1999] Iivari, J., and Linger, H. (1999). Knowledge work as collaborative work: A situated activity theory view. Proceedings of the Hawaiian International Conference on Systems Science (HICSS'32).

[Jardine & Van Rijsbergen, 1971] Jardine, N., and Van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. Information Storage and Retrieval, 7, pp. 217–240.

[Jovanović et al., 2007] Jovanović, J., Gašević, D., Knight, C., and Richards, G. (2007). Ontologies for effective use of context in e-learning settings. Educ. Technol. Soc. 10, pp. 47–59.

[Kauppinen et al., 2008] Kauppinen, T., Väätäinen, J., and Hyvönen, E. (2008). Creating and Using Geospatial Ontology Time Series in a Semantic Cultural Heritage Portal. Proceedings of the 5th European Semantic Web conference on The semantic web: research and applications (ESWC'08), pp. 110-123. Springer: Heidelberg.

[Khrennikov, 2010] Khrennikov, A. Y. (2010). Ubiquitous quantum structure: from psychology to finance. Springer: Heidelberg.

[Klarman et al., 2008] Klarman, S., Hoekstra, R., Bron, M., & others (2008). Versions and applicability of concept definitions in legal ontologies. Proceedings of OWL: Experiences and Directions (OWLED 2008 DC), Washington, DC (metro).

[Klein & Fensel, 2001] Klein, M. C., and Fensel, D. (2001). Ontology versioning on the Semantic Web. In SWWS (pp. 75–91).

[Knublauch et al., 2011] Knublauch, H., Hendler, J. A., and Idehen, K. (2011). SPIN - overview and motivation. World Wide Web Consortium, W3C Member Submission, available online: https://www.w3.org/Submission/spin-overview/

[Kontopoulos et al., 2016] Kontopoulos, E., Riga, M., Mitzias, P., Andreadis, S., Stavropoulos, T.G., Lagos, N., Vion-Dury, J.Y., Meditskos, G., Falcão, P., Laurenson, P. and Kompatsiaris, I. (2016). Ontology-based Representation of Context of Use in Digital Preservation. Accepted for publication in 1st Workshop on Humanities in the Semantic Web (WHiSe 2016) co-located with the 13th Extended Semantic Web Conference (ESWC 2016) - Heraklion, Crete, Greece, May 29, 2016.

[Koç et al., 2014] Koç, H., Hennig, E., Jastram, S., and Starke, C. (2014). State of the Art in Context Modelling–A Systematic Literature Review. In Advanced Information Systems Engineering Workshops, pp. 53-64, Springer International Publishing.

[Lagos & Vion-Dury, 2016] Lagos, N., and Vion-Dury, J-Y. (2016). Digital Preservation Based on Contextualized Dependencies. Submitted to: DocEng 2016 (on 16-03-21).

[Lagos et al., 2016] Lagos, N., Riga, M., Mitzias, P., Vion-Dury, J.Y., Kontopoulos, E., Waddington, S., Meditskos, G., Laurenson, P. and Kompatsiaris, I. (2016). Dependency Modelling for Inconsistency Management in Digital Preservation - The PERICLES Approach. Submitted to: Information Systems Frontiers Journal (on 16-02-15).

[Lötsch & Ultsch, 2014] Lötsch, J., and Ultsch, A. (2014). Exploiting the structures of the U-matrix. In Villmann, T., Schleif, F-M., Kaden, M., Lange, M. (Eds.): Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of WSOM-2014, Mittweida, Germany, July 2-4, pp. 249-258. Springer: New York.

[Luhn, 1957] Luhn, H.P. (1957). A statistical approach to mechanised encoding and searching of library information. IBM Journal of Research and Development, 1, pp. 309-317.

[Luhn, 1960] Luhn, H.P (1960). Keyword-in-context index for technical literature. American Documentation 11(4), pp. 288-295.

[Maillot & Thonnat, M. 2005] Maillot, N., and Thonnat, M. (2005). A Weakly Supervised Approach for Semantic Image Indexing and Retrieval. In CIVR, pp. 629-638.

[McAuley & Leskovec, 2013] McAuley, J., and Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. Proc. 7th ACM Conference on Recommender Systems (RecSys-13), pp. 165-172. ACM.

[Meroño-Peñuela et al., 2013] Meroño-Peñuela, A., Guéret, C., Hoekstra, R., and Schlobach, S. (2013). Detecting and reporting extensional concept drift in statistical linked data. In Proceedings of the 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC 2013.

[Mettouris & Papadopoulos, 2013] Mettouris, C. and Papadopoulos, G. A. (2013). Contextual modelling in context-aware recommender systems: a generic approach. In Web Information Systems Engineering–WISE 2011 and 2012 Workshops (pp. 41-52). Springer Berlin Heidelberg.

[Milgram, 1967] Milgram, S. (1967). Systematic study of the “small-world problem”. Psychology Today, Vol 1, No. 1, pp. 61-67.

[Mitzias et al., 2015] Mitzias, P., Riga, M., Waddington, S., Kontopoulos, E., Meditskos, G., Laurenson, P., and Kompatsiaris, I. (2015). An Ontology Design Pattern for Digital Video. Proc. 6th Workshop on Ontology and Semantic Web Patterns (WOP 2015) co-located with 14th Int. Semantic Web Conf. (ISWC 2015), CEUR-WS Vol-1461, Bethlehem, Pennsylvania, USA, October 11-15.

[Monge & Elkan, 1996] Monge A. E., and Elkan C. (1996). The Field Matching Problem: Algorithms and Applications. In KDD, pp. 267–270.

[Moore, 2008] Moore, R. (2008). Towards a Theory of Digital Preservation. The International Journal of Digital Curation, 1(3), pp. 63-75.

[Oberle et al., 2007] Oberle, D., Ankolekar, A., Hitzler, P., Cimiano, P., Sintek, M., Kiesel, M., Mougouie, B., Baumann, S., Vembu, S., and Romanelli, M. (2007). DOLCE ergo SUMO: On Foundational and Domain Models in the SmartWeb Integrated Ontology (SWIntO). J. Web Sem. 5(3), pp. 156-174.

[Olsen et al., 1993] Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B., and Williams, J.G. (1993). Visualization of a document collection: the VIBE system. Information Processing & Management 29(1), pp. 69-81.

[Ou et al., 2006] Ou, S., Georgalas, N., Azmoodeh, M., Yang, K., and Sun, X. (2006). A model driven integration architecture for ontology-based context modelling and context-aware application development, in: Model Driven Architecture–Foundations and Applications. Springer, pp. 188–197.

[Padó & Lapata, 2007] Padó, S., and Lapata, M. (2007). Dependency-based construction of semantic space models. Computational Linguistics 33(2), pp. 161-199.

[Pareti et al., 2015] Pareti, P., Klein, E., and Barker, A. D. (2015). A Linked Data scalability challenge: concept reuse leads to semantic decay. In WebSci'15 ACM Web Science Conference. ACM Press-Association for Computing Machinery.

[PERICLES D2.3.1, 2014] PERICLES Consortium, Deliverable 2.3.1: Media and Science Case Study Functional Requirements and User Descriptions, June 2014.

[PERICLES D2.3.2, 2015] PERICLES Consortium, Deliverable 2.3.2: Data Surveys and Domain Ontologies, September 2015.

[PERICLES D3.2, 2014] PERICLES Consortium, Deliverable 3.2: Linked Resource Model, July 2014.

[PERICLES D3.3, 2015] PERICLES Consortium, Deliverable 3.3: Semantics for Change Management, July 2015.

[PERICLES D4.1, 2014] PERICLES Consortium, Deliverable 4.1: Initial Version of Environment Information Extraction Tools, September 2014.

[PERICLES D4.3, 2016] PERICLES Consortium, Deliverable 4.3: Content Semantics and Use Context Analysis Techniques, January 2016.

[PERICLES D5.2, 2015] PERICLES Consortium, Deliverable 5.2: Basic tools for Digital Ecosystem Management, October 2015.

[Pruitt & Grudin, 2003] Pruitt, J., and Grudin, J. (2003). Personas: practice and theory. In Proceedings of the 2003 conference on Designing for user experiences (DUX '03). ACM, New York, USA, pp. 1-15.

[Rada et al., 1989] Rada, R., Mili, H., Bichnell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics, pp 17-30.

[Ranganathan et al., 2003] Ranganathan, A., McGrath, R.E., Campbell, R.H., and Mickunas, M.D. (2003). Ontologies in a pervasive computing environment, in: Workshop on Ontologies in Distributed Systems at IJCAI, Acapulco, Mexico. Citeseer.

[Raubal, 2008] Raubal, M. (2008). Representing Concepts in Time. In C. Freksa et al. (Eds.), Spatial Cognition VI, LNAI 5248, pp. 328-343. Springer: Berlin.

[Rayson, 2008] Rayson, P. (2008). From keywords to key semantic domains. International Journal of Corpus Linguistics. 13:4 pp. 519-549. DOI: 10.1075/ijcl.13.4.06ray

[Ren, 2014] Ren, F. (2014). Learning time-sensitive domain ontology from scientific papers with a hybrid learning method. Journal of Information Science 40(3), pp. 329–345.

[Rice, 2015] Rice, D. (2015). Sustaining Consistent Video Presentation. Tate Research Articles, available online: http://www.tate.org.uk/research/publications/sustaining-consistent-video-presentation

[Robertson & Zaragoza, 2009] Robertson, S., and Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval 3 (4), pp. 333-389.

[Robertson & Spärck Jones, 1976] Robertson, S.E., and Spärck Jones, K. (1976). Relevance weighting of search terms. JASIS, 27 (3), pp. 129-146.

[Sadrzadeh & Grefenstette, 2011] Sadrzadeh, M., and Grefenstette, E. (2011). A Compositional Distributional Semantics, Two Concrete Constructions, and Some Experimental Evaluations. In D. Song et al. (Eds.): Proceedings of QI 2011, pp. 35-47. Springer: Berlin.

[Schlieder, 2010] Schlieder, C. 2010. Digital heritage: Semantic challenges of long-term preservation. Semantic Web (1)1-2, pp. 143-147.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal 27 (July and October), pp. 379–423, 623–656.

[Sheng & Benatallah, 2005] Sheng, Q.Z., and Benatallah, B. (2005). ContextUML: a UML-based modeling language for model-driven development of context-aware web services, in: Mobile Business, 2005. ICMB 2005. International Conference on. IEEE, pp. 206-212.

[Simons & Wirtz, 2007] Simons, C., and Wirtz, G. (2007). Modeling context in mobile distributed systems with the UML. J. Vis. Lang. Comput. 18, pp. 420-439.

[Simou et al., 2005] Simou, N., Saathoff, C., Dasiopoulou, S., Spyrou, E., Voisine, N., Tzouvaras, V., Kompatsiaris, I., Avrithis, Y., and Staab, S. (2005). An Ontology Infrastructure for Multimedia Reasoning. In: Proc. International Workshop on Very Low Bitrate Video Coding (VLBV 2005), Sardinia, Italy.

[Socher et al., 2012] Socher, R., Huval, B., Manning, C.D., Ng, A.Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In Proceedings of EMNLP-CoNLL-12, Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1201-1211.

[Stojanovic et al., 2002] Stojanovic, L., Maedche, A., Motik, B., and Stojanovic, N. (2002). User-driven ontology evolution management. In Knowledge engineering and knowledge management: ontologies and the semantic web (pp. 285–300). Springer.

[Strang & Linnhoff-Popien, 2004] Strang, T., and Linnhoff-Popien, C. (2004). A context modeling survey, in: Workshop Proceedings.

[Strimpakou et al., 2005] Strimpakou, M., Roussaki, I., Pils, C., Angermann, M., Robertson, P., and Anagnostou, M. (2005). Context modelling and management in ambient-aware pervasive environments, in: Location-and Context-Awareness. Springer, pp. 2–15.

[Suárez-Figueroa et al., 2009] Suárez-Figueroa, M., Gómez-Pérez, A., and Villazón-Terrazas, B. (2009). How to write and use the Ontology Requirements Specification Document, Proceedings of the 8th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE 2009), Vilamoura, Algarve-Portugal.

[Suárez-Figueroa et al., 2012] Suárez-Figueroa, M., Gómez-Pérez, A., Motta, E., and Gangemi, A. (2012). The NeOn Methodology for Ontology Engineering. In Ontology Engineering in a Networked World, Springer Berlin Heidelberg, pp. 9-34.

[Tosi et al., 2014] Tosi, A., Olier, I., and Vellido, A. (2014). Probability ridges and distortion flows: Visualizing multivariate time series using a variational Bayesian manifold learning method. In Villmann, T., Schleif, F-M., Kaden, M., Lange, M. (Eds.): Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of WSOM-2014, Mittweida, Germany, July 2-4. pp. 55-64. Springer: New York.

[Trier, 1934] Trier, J. 1934. Das sprachliche Feld. Neue Jahrbücher für Wissenschaft und Jugendbildung, 10, pp. 428–449.

[Tsinaraki & Christodoulakis, 2007] Tsinaraki, C., and Christodoulakis, S. (2007). Interoperability of XML Schema Applications with OWL Domain Knowledge and Semantic Web Tools. On the Move to Meaningful Internet Systems (OTM), Confederated International Conferences, Vilamoura, Portugal, pp. 850-869.

[Tsinaraki et al., 2007] Tsinaraki, C., Polydoros, P., and Christodoulakis, S. (2007). Interoperability Support between MPEG-7/21 and OWL in DS-MIRF. IEEE Trans. Knowl. Data Eng. 19(2), pp. 219-232.

[Tury & Bieliková, 2006] Tury, M., & Bieliková, M. (2006, July). An approach to detection ontology changes. In Workshop proceedings of the sixth international conference on Web engineering (p. 14). ACM.

[Uexküll & Kriszat, 1956] Uexküll, J.J. and Kriszat G. (1956). Streifzüge durch die Umwelten von Tieren und Menschen: Ein Bilderbuch unsichtbarer Welten. Bedeutungslehre. Rowohlt.

[Ultsch, 2005] Ultsch, A. (2005). Clustering with SOM: U* C. In Proceedings of the 5th Workshop on Self-Organizing Maps, Vol. 2, pp. 75-82.

[Uschold, 2000] Uschold, M. (2000). Creating, integrating and maintaining local and global ontologies. In Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI-2000). Citeseer.

[Van den Bergh & Coninx, 2006] Van den Bergh, J., and Coninx, K. (2006). Cup 2.0: High-level modeling of context-sensitive interactive applications, in: Model Driven Engineering Languages and Systems. Springer, pp. 140–154.

[Vardigan & Whiteman, 2007] Vardigan, M., and Whiteman, C. (2007). ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. Archival Science, 7(1), pp. 73-87.

[Veltman, 1996] Veltman, F. 1996. Defaults in update semantics. Journal of Philosophical Logic 25(3), pp. 221-261.

[Vembue et al., 2006] Vembue, S., Kiesel, M., Sintek, M., and Bauman, S. (2006). Towards Bridging the Semantic Gap in Multimedia Annotation and Retrieval. Proc. Workshop on Semantic Web Annotations for Multimedia (SWAMM), Edinburgh, Scotland.

[Vorontsov et al., 2015a] Vorontsov, K., Frei, O., Apishev, M., Romov, P., and Dudarenko, M. (2015). BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. In Proceedings of AIST 2015, CCIS 542, Springer, pp. 370-381.

[Vorontsov et al., 2015b] Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., & Yanina, A. (2015, October). Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. In Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications (pp. 29-37). ACM.

[Vorontsov & Potapenko, 2015] Vorontsov, K., & Potapenko, A. (2015). Additive regularization of topic models. Machine Learning, 101(1-3), 303-323.

[W3C, 2012] OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012, available online: http://www.w3.org/TR/owl2-overview/

[Wang et al., 2009] Wang, S., Schlobach, S., Takens, J., and Van, W. (2009). Mapping-chains for studying concept shift in political ontologies. Ontology Matching, 13.

[Wang et al., 2011] Wang, S., Schlobach, S., and Klein, M. (2011). Concept drift and how to identify it. Web Semantics: Science, Services and Agents on the World Wide Web, 9(3), pp. 247–265.

[Wang & McCallum, 2006] Wang, X., & McCallum, A. (2006, August). Topics over time: a non-Markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 424-433). ACM.

[Widdows, 2004] Widdows, D. (2004). Geometry and meaning. CSLI Publications: Stanford.

[Widdows & Cohen, 2016] Widdows, D., and Cohen, T. 2016. Graded Semantic Vectors: An Approach to Representing Graded Quantities in Generalized Quantum Models. In Proceedings of Quantum Interaction-15. LNCS 9535, 231-244.

[Wise, 1999] Wise, J.A. 1999. The Ecological Approach to Text Visualization. Journal of the American Society for Information Science 50(13), 1224–1233.

[Wittek & Darányi, 2011] Wittek, P., and Darányi, S. (2011). Spectral Composition of Semantic Spaces. Proceedings of QI-11.

[Wittek et al., 2013] Wittek, P., Koopman, B., Zuccon, G., and Darányi, S. (2013). Combining Word Semantics within Complex Hilbert Space for Information Retrieval. Proceedings of QI-13.

[Wittek et al., 2015a] Wittek, P., Gao, S. C., Lim, I. S., and Zhao, L. (2015). Somoclu: An efficient parallel library for self-organizing maps. arXiv:1305.1422.

[Wittek et al., 2015b] Wittek, P., Darányi, S., Kontopoulos, E., Moysiadis, T., and Kompatsiaris, I. (2015). Monitoring term drift based on semantic consistency in an evolving vector field. In Neural Networks (IJCNN), 2015 International Joint Conference on (pp. 1–8). IEEE.

[Wittgenstein, 1963] Wittgenstein, L. 1963. Philosophical investigations. Blackwell: Oxford. 43.

[Wu & Palmer, 1994] Wu, Z., and Palmer, M. (1994). Verb semantics and lexical selection. Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics, pp 133-138.

[Yan et al., 2013] Yan, X., Guo, J., Liu, S., Cheng, X., & Wang, Y. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In Proceedings of the SIAM International Conference on Data Mining.

[Yildiz, 2006] Yildiz, B. (2006). Ontology evolution and versioning. Vienna University of Technology, Karlsplatz.

[Zhang & Wang, 2005] Zhang, D., Gu, T., and Wang, X., 2005. Enabling context-aware smart home with semantic web technologies. Int. J. Hum.-Friendly Welf. Robot. Syst. 6, pp. 12–20.

[Zhang et al., 2011] Zhang, S., McCullagh, P., Nugent, C., Zheng, H., and Black, N. (2011). An Ontological Approach for Context-Aware Reminders in Assisted Living' Behavior Simulation. In J. Cabestany, I. Rojas, and G. Joya (Eds.), Advances in Computational Intelligence (pp. 677–684). Springer Berlin Heidelberg.

[Zuo & Xu, 2014] Zuo, Y., Zhao, J., & Xu, K. (2014). Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems, 1-20.